Crowd-Assisted Annotation of Classical Music Compositions

Samiotis, I.P.

DOI
10.4233/uuid:33295c07-041b-4a62-953a-8a4fb4a12a03

Publication date
2024

Document Version
Final published version

Citation (APA)
Samiotis, I. P. (2024). *Crowd-Assisted Annotation of Classical Music Compositions*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:33295c07-041b-4a62-953a-8a4fb4a12a03

Crowd-Assisted Annotation of Classical Music Compositions

Crowd-Assisted Annotation of Classical Music Compositions

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op Woensdag 11 December 2024 om 15.00 uur

door

Ioannis Petros SAMIOTIS

Master of Science in Computer Science
Delft University of Technology, The Netherlands
Master of Science in Digital Visual Effects
University of Kent, United Kingdom
born in Athens, Greece.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. ir. G.J.P.M. Houben
promotor: Prof. dr. ir. A. Bozzon
copromotoren: Dr. C. Lofi

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. G.J.P.M. Houben	Delft University of Technology, promotor
Prof.dr.ir. A. Bozzon	Delft University of Technology, promotor
Dr. C. Lofi	Delft University of Technology, copromotor
Onafhankelijke leden:	
Dr. E. Law	University of Waterloo, Canada
Dr. J. Calvo-Zaragoza	University of Alicante, Spain
Prof.dr. A. Volk	University of Utrecht
Prof.dr. A. Hanjalic	Delft University of Technology
Prof.dr. G.W. Kortuem	Delft University of Technology, reserve member

SIKS Dissertation Series No. 2024-48
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems. This work is partially supported by the European Commission under the TROMPA project (H2020 770376).

Keywords: Music Annotation, Crowdsourcing, Human-in-the-loop, Music Perception, Classical Music, Hybrid Annotation Workflows
E-mail: ipsamiotis@gmail.com

ISBN: 978-94-6366-973-3

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

I am made and remade continually.

Virginia Woolf

ACKNOWLEDGEMENTS

Four years of research condensed into a book; six in total including the later work and refinement to bring this book to life. It was a period filled with events, explorations, and people that defined my interests and who I currently am in many ways. While this thesis represents a distillation of my work with people and hybrid annotation workflows, it omits to represent a big part of what it means to pursue a PhD. All the research paths that led to dead ends, the countless hours of discussions with fellow researchers and study participants, and of course all the studies that didn't end up published. And although you will find it only briefly mentioned throughout the sections, the COVID-19 pandemic had a great impact on the types of studies we could conduct during this PhD. Quarantine measures meant that we had to drop a lot of our previous plans involving live participants in music events and workshops. It also meant that we didn't pursue any in-person brainstorming with music enthusiasts and experts for the safety of everyone involved.

Nevertheless, people have been the main theme of this thesis. From their perception of music to their feedback on online tasks and the online interviews and discussions with experts and enthusiasts. This thesis was also part of a large European project, meaning that I had the great opportunity to contact and exchange ideas with people in a variety of research fields, all with a common interest: music.

Although a PhD trajectory can be a very personal experience that changes the candidate as a person, it is never pursued in isolation. There are many people I would like to thank and acknowledge their contributions, including those who directly contributed to a study or indirectly assisted me in my efforts throughout my doctorate work; as well as those who influenced in many ways the type of research I followed.

Starting from the people whose input and collaboration formed this thesis: Alessandro Bozzon and Christoph Lofi, my PhD supervisors. Their feedback and guidance showed me how to approach research systematically, focusing on the core of the matters. I would like to also thank Geert-Jan Houben, who oversaw my work and trajectory, where his feedback and discussions always helped me see the bigger picture when conducting research.

Within the Web Information Systems group –where I worked throughout my PhD– I had the chance to be colleagues with people who I highly value and some who were fundamental to my research. I would like to start with Andrea Mauri, who was the first person I met and interacted with within the group, and was destined to be a great mentor, collaborator, and close friend. He contributed to the majority of the studies presented here (directly or otherwise) and always entertained my out-of-the-box ideas with care and enthusiasm. I would like to also thank Sihang Qiu, another great colleague and friend who significantly helped me navigate PhD, especially in its beginnings. His approach to experimental setup highly influenced my current approaches. On great colleagues within the group, I am grateful to Achilleas Psyllidis for helping me overcome roadblocks in my research and Ujwal Gadiraju for broadening my horizons on human computation. Special thanks also go to Agathe Balayn and Christos Koutras, with whom I frequently exchanged academic wisdom. Finally, I would like to generally thank my office mates Sepideh, Shabnam, Shahin, Kyriakos, and Giorgos, as well as my fellow colleagues at the time: Oana, Tim, Manuel, Felipe, Ziyu, David, Nirmal, Arthur, Guz, Sara, Andra, Alisa, Peide, Garrett, and Lorenzo.

Fundamental parts of my PhD were also my interactions with the Multimedia Computing group. Cynthia Liem has been an invaluable mentor already since my MSc in TU Delft, and my collaboration with her throughout my PhD was critical for both conducting research on music information retrieval and collaborating with partners within the TROMPA project. The mentorship and general help I received from the group extended also through Jaehun Kim, who I met while he supervised me during my MSc thesis and with whom I am grateful to share a unique bond. Additionally, I would like to mention Julián, Andrew, and Sandy, with whom I had numerous discussions that shaped my PhD in many ways.

I would like to also thank collaborators within the TROMPA project. In its duration, I met people who inspired me both in an academic context but also in aspects of music tech and digitization. I would like to especially mention Marcel van Tilburg for the very thought-provoking discussions and the valuable input he gave me regarding professional orchestras and their technological needs. Tackling those technological needs and the complex engineering challenges of TROMPA deliverables would have been impossible without the invaluable help of Carlos and Shaad. Their engineering prowess was pivotal for launching Scriptoria live.

As mentioned in the beginning, this thesis took an extra two years to be finalized. This gave me a unique opportunity to see my thesis' content, its scope, and reach with a fresh pair of eyes. I was able to discuss it with people in the field who were familiar with the digitization challenges of classical music or manually annotating music in general. I would like to thank here Minna Ylikauma who dedicated time to discuss with me her experiences with classical music metadata and the applicability of methods beyond the genre, which shaped my approach to this thesis' conclusions. I should also mention here Paul Natale, who put me in touch with Minna and pulled the curtain on how music metadata are generated and handled in the industry. While mentioning Paul, I would like to also thank David and Anders, my other fellow bandmates in Rosa Canina, for all the fun moments we spent throughout rehearsals, recordings, and gigs, giving me a much-needed mental boost for the final miles in my thesis.

Emotional support came in many shapes and forms throughout my PhD trajectory. My partner, friends, and family all played a crucial role in supporting me, and for that, I'm thankful to each one of them. Coffees, dinners, and general discussions I shared with Pantelis, Inger-Lise, Aparajita, Alyssa, Vishruty, Anurag, Valentina, Salil, Aksel, Amko, Michiel, Amrita, Anne, Jesse, Jeff, Pascal, Celine, Yasin, Alex, Jini, Yen-Chia, all played their role in getting this thesis in the form that is presented here. Special mention, to a group of friends, Andrea, Jay, Vera, Sandy, and Pedro, with whom I shared an epic storytelling journey together, starting from the beginning of my PhD, and sharing unique memories ever since.

The support I received from my partner, parents, and brother in times of need was pivotal, while their influence on this PhD is apparent to those who know my roots. Being raised in a family of theater actors, special effects supervisors and photographers played a big role in my pursuit of this thesis' topic. Their encouragement from a young age to pick up and study music led me to some of the most memorable moments in my life. Unfortunately, I lost my father right after my first year of PhD, but my memories of him helped me throughout its duration and will still carry me throughout life.

Finally, I would like to thank my life partner Francesca, where we both met in the beginnings of our PhD trajectories and we have been there for each other ever since. She made the rough moments of this 6-year period smoother, while amplifying the happy ones. And while this chapter ends, I couldn't be more excited about the next one; where I'm already getting perseverance lessons from our little one.

Petros
Rotterdam, 2024

TERMINOLOGIES

Term	Definition
Classical Music	"Classical music" is the music genre that encompasses musical traditions and stylistic expressions that belong under the category of "Western Classical Music".
Composer	A person who writes music, often as a profession. It is mainly used in the context of classical music in this thesis.
Composition	A musical work created by a composer. It is mainly used in the context of classical music in this thesis.
Music Score	The physical or digital form of the notation-based representation of a composition.
Sheet Music	The physical form of the notation-based representation of a composition–a physical music score.
Optical Music Recognition (OMR)	The research field that investigates how the computer can read music notations on sheet music.
Digitization	The process of transforming the physical form of an artifact into its digital representation.
Crowdsourcing	The practice of gathering annotations and information from users on online platforms through tasks designed for the purpose.
Nichesourcing	Crowdsourcing workflows that incorporate a narrower set of annotators (*"niche" groups*), often specialized on specific knowledge domains.
Human Computation	The research field dedicated to understanding how to incorporate people in computational methods that are hard to automate while also studying their computational properties and characteristics.
Human-in-the-loop	Data processing pipelines that incorporate human feedback alongside automated methods.
Transcription (Process)	*Transcription, digital transcription* or *music transcription* of sheet music is defined as the process through which a scanned music score is converted into a machine-readable format. Transcription processes can include both automated and crowdsourced tasks.
Transcription (Artifact)	*Transcription(s), digital transcription(s)* is/are the product of transcription processes, where sheet music has been digitally transformed into a machine-readable format.
Transcription task	A crowdsourcing task where users are asked to contribute to the transcription processes by detecting, recognizing, or transcribing specific elements on a music score.
(Music) Annotations	The term *(music) annotations* is used as an umbrella term to differentiate products of transcription processes with annotations that include but are not limited to links to other classical music data entities, contextual annotations, and recording-specific annotations.
OMR System	A computational system that implements OMR methods to detect, recognize and transcribe sheet music to its digital version.

Table 1: Summary of key terms and corresponding definitions.

SUMMARIES

ENGLISH

Music annotation and transcription of music sheets are traditionally performed by experts. Although these processes result in high quality data, the scope of each effort is relatively narrow resulting in highly specialised and specific datasets of annotated music compositions, which leads to a fragmentation in the design efforts for automated tools. In music traditions such as classical music, the shortcomings of current digitization workflows become even clearer: due to the vast corpus and varying stylistic intricacies, experts tend to have specific knowledge and take up projects that concern very specific periods or composers, limiting our reach in regards to conserving classical music information as a whole.

On the other hand, crowdsourcing has been successfully utilized in other domains for annotating different modalities (text, image, video, audio), despite the unreliable pool of expertise on online platforms. Commercially successful projects have utilized the crowd, which provided annotations of adequate quality. These annotations were later used to fuel machine learning methods that rely on bulks of annotated data to perform automatic classification, regression, and detection tasks. However, due to the complexity of music as an artifact, there are still only a few examples where the crowd was integrated into any form outside of subjective annotation tasks (e.g., indicate the mood of the excerpt).

In this thesis, we tackle this research gap of integrating the crowd in the annotation processes of music compositions. We surveyed current practices on Optical Music Recognition to identify parts where the crowd could assist, alongside proposing hybrid annotation workflows for music compositions. We studied the capabilities of online participants with unknown musical expertise, quantified their musical abilities, and related them to their performance in music annotation tasks. With our goal being to identify ways to expand the preservation efforts for classical music through the assistance of the general public, we investigated potential online sources of music information and prospective participants outside the currently available crowdsourcing platforms. To that end, we studied how composers' popularity manifests on community-driven platforms through the interactions of music enthusiasts online. We also conducted interviews and focus group discussions with experts and semi-experts, to understand their quality requirements on semantically-rich digital music scores, and identified transcription patterns that could inform our task designs. We finally delivered our system architecture, which combines computer vision and algorithmic scheduling, with microtasks designed to be performed by human annotators in parallel.

Our findings show that with the right methods to quantify the musical competence of a person, paired with careful design of the annotation tasks and interfaces, we can successfully integrate the crowd in the music annotation processes, to generate meaningful and useful information regarding classical music compositions and beyond. This

thesis enables future research by showcasing the versatility of the crowd and providing task design methods to accommodate their lack of formal training in the field. It also provides experimental methods to reliably identify how different music composition elements affect the crowd's performance, alongside proposing user interface elements that can mediate the complexity of the artifacts. Practices such as those presented in this thesis can lead to scaling up our digitization efforts, generating accurate and useful annotations through the crowd, even in such domain-specific and knowledge-intensive topics as classical music compositions.

NEDERLANDS

Muziekannotatie en -transcriptie worden traditioneel uitgevoerd door experts. Hoewel deze processen muziekannotaties van hoge kwaliteit geven, is de reikwijdte van elke inspanning relatief klein, wat resulteert in zeer gespecialiseerde datasets van geannoteerde muziekcomposities. In muziektradities zoals klassieke muziek worden de tekortkomingen van de huidige workflows nog duidelijker: door het enorme corpus en de verschillende stilistische fijne kneepjes hebben experts de neiging om specifieke kennis in te brengen en projecten op te nemen die zeer specifieke periodes of componisten betreffen.

Aan de andere kant is crowdsourcing met succes gebruikt in andere domeinen voor het annoteren van verschillende modaliteiten (tekst, afbeelding, video, audio), ondanks de onbetrouwbare pool van expertise op online platforms. Commercieel succesvolle projecten hebben gebruik gemaakt van de menigte, die annotaties van voldoende kwaliteit leverde. Deze annotaties werden later gebruikt om machine learning-methoden mee te voeden die afhankelijk zijn van de bulk van geannoteerde gegevens om automatische classificatie-, regressie- en detectietaken uit te voeren. Maar vanwege de complexiteit van muziek als artefact zijn er nog maar weinig voorbeelden waarbij de menigte in welke vorm dan ook buiten subjectieve annotatietaken werd geïntegreerd (bijv. om de stemming van het fragment aangeven).

In dit proefschrift pakken we deze onderzoekskloof aan van het integreren van de menigte in de annotatieprocessen van muziekcomposities. We hebben de huidige praktijken op het gebied van optische muziekherkenning onderzocht om delen te identificeren waar het publiek zou kunnen helpen, en om daarnaast voorstellen te doen van hybride annotatieworkflows voor muziekcomposities. We bestudeerden de mogelijkheden van online deelnemers met onbekende muzikale expertise, kwantificeerden hun muziekvaardigheden en relateerden ze aan hun prestaties in muziekannotatietaken. Om potentiële informatiebronnen en deelnemers buiten crowdsourcingplatforms te identificeren, hebben we bestudeerd hoe de populariteit van componisten zich manifesteert op community-driven platforms door de interacties van muziekliefhebbers. We hebben ook interviews en focusgroepdiscussies gehouden met experts en semi-experts, om hun kwaliteitsvereisten voor semantisch rijke digitale muziekpartituren te begrijpen, en om transcriptiepatronen te identificeren die onze taakontwerpen zouden kunnen informeren. We hebben tenslotte onze eigen systeemarchitectuur geleverd die computervisie en algoritmische planning combineert, met microtaken die zijn ontworpen om parallel door menselijke annotatoren te worden uitgevoerd.

Onze bevindingen laten zien dat met de juiste methoden om de muzikale competentie van een persoon te kwantificeren, gecombineerd met een zorgvuldig ontwerp van de annotatietaken en -interfaces, we het publiek met succes kunnen integreren in de muziekannotatieprocessen. Dit proefschrift maakt toekomstig onderzoek mogelijk door de veelzijdigheid van de menigte te laten zien en taakontwerpmethoden aan te bieden om tegemoet te komen aan het gebrek aan formele training in het veld. De thesis biedt ook experimentele methoden om op betrouwbare wijze te identificeren hoe verschillende elementen van muziekcomposities de prestaties van het publiek beïnvloeden, en doet voorstellen van gebruikersinterface-elementen die met de complexiteit van de artefacten kunnen omgaan. Praktijken zoals die in dit proefschrift worden gepresenteerd, kunnen leiden tot het opschalen van onze digitaliseringsinspanningen en het genereren van nauwkeurige en nuttige annotaties door de menigte, zelfs in zo'n domeinspecifiek en kennisintensief onderwerp als klassieke muziekcomposities.

Ελληνικά

Η επισημείωση και μεταγραφή της μουσικής εκτελούνται, ως επί το πλείστον, από ειδικούς, μέσω διαδικασιών που οδηγούν στην παραγωγή μουσικών δεδομένων υψηλής ποιότητας. Ωστόσο το εύρος κάθε τέτοιας προσπάθειας είναι σχετικά περιορισμένο, με αποτέλεσμα την δημιουργία εξαιρετικά εξειδικευμένων συνόλων μουσικών δεδομένων και στην επισημείωση συγκεκριμένων μόνο μουσικών συνθέσεων. Σε μουσικά είδη όπως η κλασική μουσική, τα μειονεκτήματα των υπαρχόντων προσπαθειών είναι ακόμη πιο εμφανή: λόγω του τεράστιου όγκου μουσικών συνθέσεων και των ποικίλων στυλιστικών ιδιαιτεροτήτων, οι ειδικοί τείνουν να έχουν εξειδικευμένες γνώσεις και να απασχολούνται με εγχειρήματα που αφορούν πολύ συγκεκριμένες περιόδους ή συνθέτες.

Αφετέρου, παρόλο που η τεχνογνωσία είναι δυσεύρετη στις διαδικτυακές πλατφόρμες, ο πληθοπορισμός (crowdsourcing) έχει χρησιμοποιηθεί με επιτυχία σε άλλους τομείς δημιουργίας δεδομένων ποικίλων μορφών (κείμενο, εικόνα, βίντεο, ήχος). Μέχρι τώρα, πολλά εμπορικώς επιτυχημένα εγχειρήματα έχουν χρησιμοποιήσει το διαδικτυακό πλήθος (crowd) για την παροχή δεδομένων και επισημειώσεων επαρκούς ποιότητας. Αυτές οι επισημειώσεις έχουν χρησιμοποιηθεί για να τροφοδοτήσουν μεθόδους Μηχανικής Εκμάθησης (Machine Learning) οι οποίες βασίζονται σε μεγάλους όγκους επισημειωμένων δεδομένων για την εκτέλεση εργασιών αυτόματης ταξινόμησης (classification), ανάλυσης παλινδρόμησης (regression analysis) και ανίχνευσης (detection). Ωστόσο, λόγω της πολυπλοκότητας της κλασσικής μουσικής, υπάρχουν παρά μόνο λίγα παραδείγματα που το διαδικτυακό πλήθος ενσωματώθηκε επιτυχώς σε κάποια μορφή διαδικασιών που ζητούν κάτι πέραν υποκειμενικών επισημειώσεων και δεδομένων (π.χ. υπόδειξη των συναισθημάτων ενός μουσικού αποσπάσματος).

Σε αυτή τη διατριβή, αντιμετωπίζουμε αυτό το ερευνητικό κενό της ένταξης του διαδικτυακού πλήθους στις διαδικασίες δημιουργίας δεδομένων και επισημειώσεων μουσικών συνθέσεων. Ερευνήσαμε τις υπάρχοντες πρακτικές σχετικά με την Οπτική Αναγνώριση Μουσικής (Optical Music Recognition) για να εντοπίσουμε διαδικασίες όπου το πλήθος θα μπορούσε να βοηθήσει. Παράλληλα, προτάσσουμε την συνύπαρξη αυτόματων διαδικασιών και ανθρώπινου δυναμικού, για την δημιουργία υβριδικών ροών εργασίας και επισημείωσης των μουσικών συνθέσεων. Μελετήσαμε τις δυνατότητες διαδικτυακών συμμετεχόντων με προηγουμένως άγνωστη μουσική εμπειρία, ποσοτικοποιήσαμε τις μουσικές τους ικανότητες και τις συσχετίσαμε με την απόδοσή τους σε εργασίες μουσικής επισημείωσης. Στην συνέχεια, για να εντοπίσουμε συμμετέχοντες εκτός των πλατφορ-

μών πληθοπορισμού (crowdsourcing) καθώς και για την εύρεση άλλων πιθανών πηγών μουσικής πληροφορίας, μελετήσαμε σε διαδικτυακές πλατφόρμες κοινοτήτων, το πώς η δημοτικότητα των συνθετών κλασσικής μουσικής εκδηλώνεται μέσω των αλληλεπιδράσεων των λάτρεων του είδους. Πραγματοποιήσαμε επίσης συνεντεύξεις και ομαδικές συζητήσεις με ειδικούς και ημι-ειδικούς, για να κατανοήσουμε τις ποιοτικές απαιτήσεις τους σε σημασιολογικά πλούσιες ψηφιακές μουσικές παρτιτούρες. Μέσω των σχολίων τους, εντοπίσαμε επίσης μοτίβα μουσικής μεταγραφής που θα μπορούσαν να ενημερώσουν τους σχεδιασμούς μας για υβριδικές ροές εργασιών. Τελικώς, παραδώσαμε και παρουσιάσαμε τη δική μας αρχιτεκτονική υβριδικού συστήματος, το οποίο συνδυάζει την μηχανική όραση (computer vision) και την αλγοριθμική χρονοδρομολόγηση μικροεργασιών, που έχουν σχεδιαστεί για να εκτελούνται παράλληλα από ανθρώπινο δυναμικό.

Τα ευρήματά μας δείχνουν ότι με τις σωστές μεθόδους ποσοτικοποίησης της μουσικής ικανότητας ενός ατόμου, και σε συνδυασμό με τον προσεκτικό σχεδιασμό των εργασιών επισημείωσης μουσικών συνθέσεων, μπορούμε να ενσωματώσουμε επιτυχώς διαδικτυακό πλήθος ποικίλης μουσικής εμπειρίας. Αυτή η διατριβή ανοίγει τις δυνατότητες για μελλοντικές έρευνες στον χώρο, προβάλλοντας την ευελιξία του διαδικτυακού πλήθους και συγχρόνως παρέχοντας μεθόδους σχεδιασμού υβριδικών εργασιών που μπορούν να καλύψουν την πιθανή έλλειψη επίσημης εκπαίδευσης στο πεδίο της κλασσικής μουσική. Παρέχει επίσης πειραματικές μεθόδους για τον αξιόπιστο προσδιορισμό του τρόπου με τον οποίο διαφορετικά στοιχεία μουσικών συνθέσεων επηρεάζουν την απόδοση αυτού του πλήθους, ενώ παράλληλα προτάσσει μεθόδους διεπαφής με τους χρήστες, ως αντίβαρο στην πολυπλοκότητα των μουσικών τεχνουργημάτων. Πρακτικές όπως αυτές, μπορούν να οδηγήσουν στην κλιμάκωση των προσπαθειών μας για ψηφιοποίηση δεδομένων πολιτισμικής κληρονομιάς, δημιουργώντας ακριβή και χρήσιμα δεδομένα μέσω του διαδικτυακού πλήθος, ακόμη και σε θέματα που χρειάζονται εξειδικευμένες γνώσεις όπως οι συνθέσεις κλασικής μουσικής.

CONTENTS

Acknowledgements **vii**

Terminology **xi**

Summaries **xiii**

1 Introduction **1**
1.1 Background and Motivation 1
1.2 The TROMPA Project 3
1.3 On Classical Music Transcriptions and Automatic Recognition 4
1.3.1 Structure and Main Components 4
1.3.2 Automatic Recognition of Music Score Elements. 6
1.3.3 Examples of Challenging Cases in Automatic Recognition 7
1.4 An Opportunity for Hybrid Annotation Workflows 9
1.5 Challenges and Research Questions. 11
1.6 Contributions and Origins of Chapters 13

2 Modeling and Measuring the Musical Affinity of Crowd Annotators **15**
2.1 Musical Properties of Annotators 16
2.1.1 Modelling Musical Competence 16
2.1.2 Sources and Versioning 18
2.2 Music Sophistication and Aural Perception Skills of the Crowd 19
2.2.1 Related Work. 20
2.2.2 Experimental Design. 21
2.2.3 Results 27
2.2.4 Discussion 38
2.2.5 Conclusion. 40
2.3 User Engagement with Classical Music on Online Platforms 41
2.3.1 Related Work. 42
2.3.2 Data Collection 43
2.3.3 Results 44
2.3.4 Discussion & Conclusion. 47

3 Designing Crowd-Enhanced Annotation Systems for Music Compositions **49**
3.1 Design, Deployment and Evaluation of Scriptoria. 50
3.1.1 System Architecture 50
3.1.2 Task Design and Interfaces. 52
3.1.3 Iterative Design and Evaluation 53
3.1.4 Results and Improvements. 56
3.1.5 Insights for Future Versions 57

3.2 Error Detection on Music Scores 59
3.2.1 Related Work. 59
3.2.2 Experimental Design. 60
3.2.3 Results 63
3.2.4 Discussion 67
3.2.5 Conclusion. 68
3.3 Temporal Activity Detection of Musical Instruments in Audio. 68
3.3.1 Related Work. 70
3.3.2 Experimental Design. 70
3.3.3 Results 73
3.3.4 Discussion 77
3.3.5 Conclusion. 78

4 Conclusions 79
4.1 Contributions and Insights 80
4.1.1 On the Musical Affinity and Competences of Non-experts 81
4.1.2 On Deploying Human-assisted Music Transcription Systems 82
4.2 Challenges and Lessons learned 83
4.3 Future Opportunities 86

Bibliography 89

List of Figures 99

List of Tables 101

Curriculum Vitæ 103

List of Publications 105

SIKS Dissertation Series 107

1

INTRODUCTION

This thesis addresses the problem of scaling up the digitization of classical music by investigating how to incorporate the general public in the annotation of music compositions. We strongly believe that the innate human understanding of music, the interest of the general public in classical music, and standardized methods to transcribe music scores (i.e., from physical form to digital, machine-readable format) can lead to a natural inclusion of non-experts in the digitization processes. In this thesis, we make use of crowdsourcing approaches to design transcription tasks and pipelines that incorporate automated and human-driven annotations, informed by analyses of the musical abilities of the crowd and novel models on annotators' expertise. Our results show the viability of scaling up the annotation and transcription processes of classical music compositions by including online, readily available annotators with diverse levels of expertise in the field.

1.1. BACKGROUND AND MOTIVATION

If an individual performs a quick online search regarding classical music, they would immediately find a sizable amount of information and artifacts about composers such as Mozart, Beethoven, and Bach, as well as their musical compositions. This abundance of information reflects the appeal that these composers have to the public, which drove, as a result, the recording, publishing, and digitization of their works throughout the years. Not all classical music information, though, has been recorded and preserved equally so far. As a matter of fact, it gradually becomes more difficult to retrieve information regarding lesser-known compositions of famous composers, with the phenomenon worsening for less popular composers.

To understand these difficulties in preservation and archiving, we need to contextualize classical music information. Within musicology, what is described as classical music can refer to a specific musical style that became popular in Western Europe during the late 18th and early 19th centuries. Colloquially though, what "classical music" often encompasses is the majority of output in the styles of Western Classical Music spanning through centuries, ever-evolving with composers from around the world outputting music material in this style. Throughout this thesis, we refer to "classical music" as the genre that encompasses Western Classical Music styles in general, letting us tackle a plethora of diverse corpora that can exhibit unique challenges in their musical pieces.

These musical works, or *compositions* as we will frequently refer to them, are the intellectual works that capture all the relevant musical elements that the composer wanted to convey, in a coherent and self-contained manner. The attention the genre has garnered is undeniable, both on pure music output (e.g. compositions and audio/visual recordings) but also academic analysis which describes and contextualises it. This has resulted in a sheer amount of ever-expanding information regarding its musical artifacts and the knowledge surrounding them.

Yet, many composers and their compositions remain obscure to the wider public. Their compositions often “live” in private collections, while few people study them and specialise in these composers, further hindering the digitization and annotation efforts in an already niche domain of expertise.

Adding to the domain complications, music compositions, as with literary works, rarely remain unchanged through time. Subsequent editions can alter the original material so that an editor, a conductor, or an instrumentalist would seem fit to their stylistic choices. This results in versions that often “sit” beside the originals, or in some cases “on top” of them, as they only exist as added notes on the previous editions. For a centuries-spanning genre, this can lead to versions of compositions having vastly different layouts or notations, often requiring further specialization of tools and proficiency.

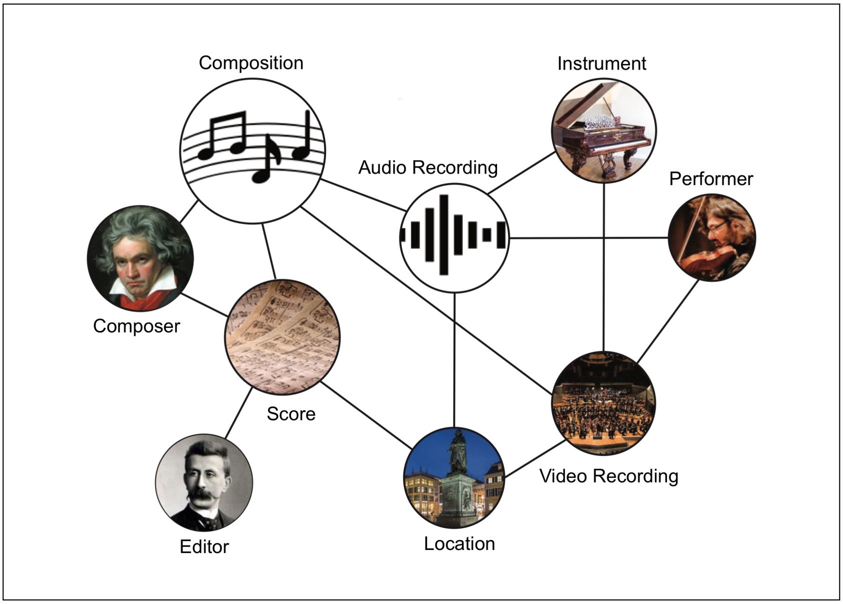


Figure 1.1: A simplified example schema of music concepts related to a classical music composition. “Composition” is the musical work of a “Composer”, published in physical media such as a “Score” (or “sheet music”). A “Score” can be edited (by an “Editor”) and re-published, while there can be multiple audio/video recordings of a musical work performed on specific “Instruments” by specific “Performers”.

To tackle such complexities and challenges, one has to be able to incorporate the different facets of classical music information and its artifacts (whether auditory, visual, or textual), which can require extensive domain knowledge. Indeed, efforts to preserve the classical music heritage are mainly performed by specialised institutions such as musical libraries and conservatories, which employ people highly specialised in terms of expertise. As expected, such experts are few and often need to be combined with other specialists who will assist in transferring physical information to a digital form (i.e., the actual contents of a music score and/or accompanying handwritten annotations). This has resulted in disconnected and highly focused data repositories, which creates a challenging landscape to develop automated computational methods that can assist our efforts in preserving classical music heritage as a whole.

We strongly believe that by designing processes that include a wider crowd (e.g., online communities, users on crowdsourcing platforms, amateur musicians), we can help expand our efforts in processing classical music information, leading to larger and richer digital repositories. These processes need to be carefully designed to be easily achievable by people with low or no musical proficiency, but to also contribute meaningful annotations that are important in automatic processing pipelines, such as semantic enrichment of music scores (e.g., explanations on phrases, connections to other compositions) and instrument recognition. That way we can democratise access to classical music information by assisting research and preservation efforts that are hindered by resource constraints, but also include interpretations, experiences, and knowledge of the general public regarding our shared cultural heritage of classical music.

1.2. The TROMPA Project

This work was partially supported by the European Commission under the Horizon H2020 project, called TROMPA (**T**owards **R**icher **O**nline **M**usic **P**ublic-domain **A**rchives). This project aimed to tackle the lack of publicly and openly available music archives, covering the different modalities of classical music information to enable future studies and research in the field.

TROMPA was an international research project involving a wide range of academic and industry partners from around Europe. Specifically, parts of this work were inspired by use cases, as discussed with the research partners at MDW (University of Music and Performing Arts Vienna), UPF (University Pompeu Fabra), Goldsmiths University of London, Concertgebouw van Amsterdam, VideoDock, and Muziekweb.

In the context of TROMPA, our goal was to research and enable methods to scale up the digital transcription of classical music. We approached this challenge by understanding the current challenges in automated methods and assessing and modeling the requirements of possible contributors in the transcription processes. We also conducted studies where we employed an untrained crowd found on online crowdsourcing platforms to evaluate the extent to which they could assist with specific transcription tasks. Although we actively engaged and assisted in design sessions and tackling engineering challenges throughout the project, we specifically produced three project deliverables on “Crowd evaluation methodologies”, “Annotator properties and metrics” and “Hybrid annotation workflows”. Towards the end of the project, we also delivered and evaluated an end-to-end, collaborative music score digitization system, which led to our contributions to another three project deliverables on “Working prototype for Orchestras” and mid-term and final project evaluations.

The delivered system would receive music scores in PDF format from the partner’s (VideoDock) graph database, segment and process the images, and finally recreate a “skeleton” of the score in MEI format[1] (Music Encoding Initiative). The segments of the score would be available on the main “campaign” web page of TROMPA project for contributors to digitize into MEI format through a web-based, fully graphical user interface. The resulting MEI was being hosted on Github[2] for versioning and making openly available the transcriptions. The system is discussed in detail in Section 1.4, and was published in the Proceedings of the 2021 Web Conference [83].

The TROMPA project and parts of our studies were quite impacted by the COVID-19 pandemic. The project’s reach to enthusiasts and musicians was limited, and in-person studies were prohibited. While engagement events that could invite a wider variety of participants in our studies were cancelled, we could still conduct studies online and even engage in online focus group discussions with semi-experts who could perform annotation tasks or benefit from the output of our online transcription system.

1.3. ON CLASSICAL MUSIC TRANSCRIPTIONS AND AUTOMATIC RECOGNITION

The preferred medium to capture the form and content of classical music has traditionally been the music scores. A music score contains the notes the instrument(s) should play, the way they should express those notes, the pauses they should take, and the speed they should be playing on. It is a way to communicate a musical idea to others through a standardized notation. As such, it is an important piece of information to preserve, as by preserving it, we conserve the musical idea that a composer wanted to communicate.

Music theory helps a composer structure their musical ideas and be able to express them through musical scores. This is a difficult task as music expresses feelings and concepts that can be hard to capture in a structured manner. As a result, several conventions have been adopted through the centuries, following the different musical traditions worldwide. Classical music is traditionally conveyed through standardized Western music notation, which contains several structural elements.

1.3.1. STRUCTURE AND MAIN COMPONENTS

To an untrained eye, a page from a classical music score can seem intimidating; there are numerous symbols that do not exist outside of the music domain, text from the Italian language, numbers in different sizes and places across the page and multiples of “lines” connecting notes in different ways.

A classical music score can contain many notations, which follow a predefined structure and are meant to capture the intricacies of the different classical music periods and “schools”. Notes are the main “words” in a music sheet, that together articulate the music passages and phrases of a piece. They dictate *what* the musician has to play on their instrument and for how long. On top of the notes and around them, there can be several

[1] An XML type of document for semantic representation of music scores: `https://music-encoding.org/`
[2] `https://github.com`

other notations that instruct on *how* to play the notes, either softly, accented, muted, and more. These are more known as "performance notations" (see yellow highlights in Figure 1.2). In this section we present a brief overview of music notations that can be found in a music score, to enable further dives on challenges and automatic methods.

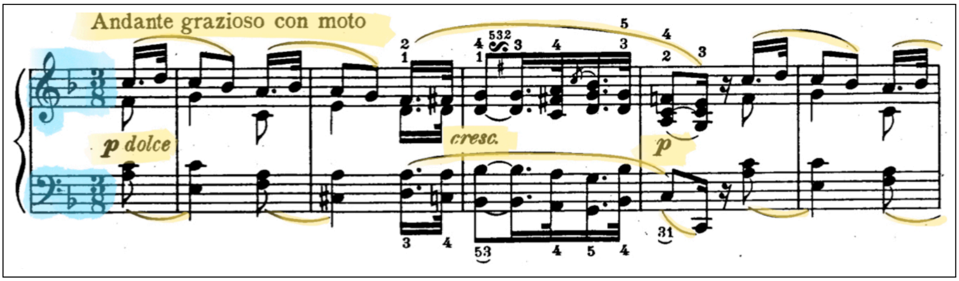


Figure 1.2: A starting system of measures in a piano piece (Andante, in F major, Ludwig van Beethoven)

While all modern scores have the 5-line structure ("staff" or "pentagram"), the pitch of the notes and the amount of notes per cell are not identical in all compositions. A "line" of staffs that contains several "cells" (measures) and extends from left to right of a page is called "system" and in the first such system of a score, we find in its beginning three of the most important notations: a "cleff", a "key signature" and a "time signature" (see blue highlights in Figure 1.2). Depending on the compositions, these three elements can change throughout the piece even within a system, between measures.

To briefly understand concepts of Western notation as they are used in classical music, the vertical position of notes on a pentagram affects their pitch, with higher positions resulting in higher registers. Cleffs dictate the names (therefore the main pitch) of the notes arranged throughout the pentagram. This means that different cleffs shift all the notes' pitches, affecting all of the notes presented after them.

The horizontal placement of notes is dictated on how someone should "count" the rhythm of the music piece. While cleffs dictate the name/pitch of notes within a pentagram, the time signature prescribes the maximum "amount of notes" that can exist within a single measure. This helps communicate how the composer intended their composition to be played, using the stems ("tail" and "flag") of notes to denote the fraction of time the note is supposed to be played for.

The vertical arrangement of notes is insufficient to communicate the wide variety of available pitches. In Western musical tradition (and classical music itself), there are semi-tones that exist "between" the space and lines of the pentagram. Those are denoted by the "accidentals", the sharp (♯), the flat (♭), double-sharp (𝄪) and double-flat (𝄫). They are meant to be used as a momentary change of a note by a semi or whole tone. When certain notes need to be considered always altered in a piece, compared to their "default" chromatic (pitch) interpretation, the composer uses a Key Signature. A key signature shows which notes should always be interpreted as a semitone higher or lower after its introduction. If a composer wants a note to be interpreted only based on its position on the pentagram, then they use the natural (♮) symbol.

Until the introduction of a different cleff, key, or time signatures, all of the notes have to be interpreted based on the preceding “values” of those notations. If the music piece dictates it, the composer can re-introduce these notations with different values which will introduce different interpretations to all the notes following them.

1.3.2. AUTOMATIC RECOGNITION OF MUSIC SCORE ELEMENTS

Optical Music Recognition (OMR) is a research field dedicated to computationally recognizing the music notations, encompassing a collection of methods and processes dedicated to automatically detect visual elements on sheet music. As previous studies suggest [76], the OMR processes can generally grouped into the following three main categories:

1. Image pre-processing
2. Music symbol segmentation and recognition
3. Semantic reconstruction

Image pre-processing. During the image pre-processing, different techniques are applied to scanned images to reduce the computational cost and make the next OMR steps more efficient. One of the most important methods of image preprocessing is “Binarization” which is the process of converting the pixel image into a binary image (black and white), separating the foreground from the background. This is a common step for most of the OMR tools. Binarization eases the OMR tasks by reducing the amount of information the following steps need to process. For example, it is easier to detect a music symbol in a binary image than in a colour image. However, binarization can also pollute the image, losing relevant details [76].

Many music scores are ancient documents in poor condition due to paper degradation (yellowing, mold, mildew, etc.), which often introduces noise to the image, reducing the quality of the OMR tool output. Therefore, working with old music score sheets requires a specialized algorithm for image-cleaning and binarization to reduce the aforementioned problems [27].

Music symbol segmentation and recognition. Music symbol segmentation is the process of locating and isolating the music object. The main objective is to find the correct position of each symbol to be identified in the next OMR step. This is one of the most challenging OMR steps and is highly error-prone. Most symbols on a music score are connected by staff lines. To isolate those symbols, staff lines must be detected and removed. Accurate staff line removal is challenging because symbols and staff lines must be disconnected without removing pixels belonging to the symbols. Unfortunately, staff lines are not always perfectly horizontal, so knowing the exact location of the staff line is required. This procedure can be even more complex due to low image quality (paper degradation, stains, etc), and zones with a high density of symbols [79].

After the segmentation stage, the segmented symbols must be recognized and classified into predefined groups, such as notes, rests, accidentals, clefs, etc. Symbol identification is a hard task because of symbol variability. Each symbol can have different variations due to different score editors or the continuous evolution of music notation over time. However, variability can also be observed within the same music score, making it even more difficult with symbol ambiguity. In addition, the previous segmentation step may have cut or degraded the objects [79].

Semantic reconstruction. The last stage of the OMR framework is to reconstruct the music semantics from the recognized symbols, combining them with the staff system to reproduce the meaning of the scanned music [76]. Unlike optical character recognition (OCR) which is predominantly one-dimensional and which can also rely on strong language pattern heuristics, OMR tools require an interpretation of two-dimensional relationships between music objects.

As a consequence, many errors may occur due to a symbol placed in the wrong position. For example, a slur symbol is a curved line generally located over the notes. If a slur is placed in a wrong position, it leads to a misinterpretation of the music score. Likewise, a dot has different meanings depending on where it is located (e.g., on top vs on the side).

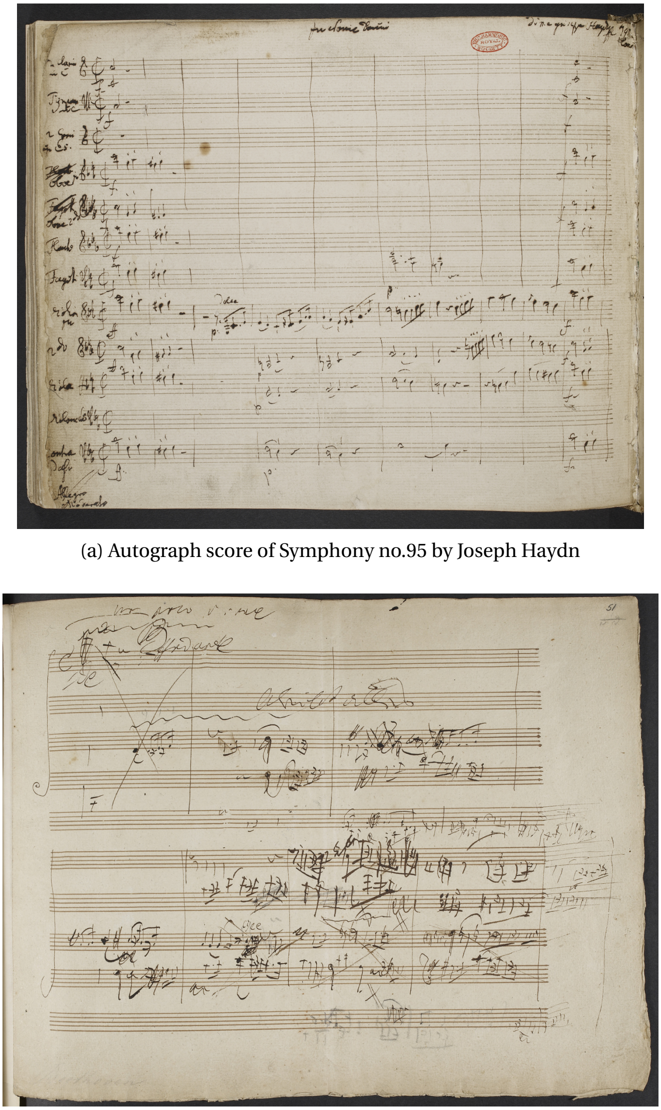
The last step of OMR systems is to export the final score into a machine-readable format. Several formats have been developed, such as MIDI (Musical Instrument Digital Interface), MusicXML, MEI (Music Encoding Initiative), NIFF (Notation Information File Format), etc. Generally, each tool has its own (set of) output formats. This lack of a commonly accepted representation imposes an obstacle for OMR tool assessment [40], further hindering any attempts for a standardized approach.

1.3.3. Examples of Challenging Cases in Automatic Recognition

Music is a complicated artifact that can be challenging to transfer to a textual format. Centuries of classical music tradition, though, have led to a condensed musical notation where every element plays an important role in how to “read” and interpret a music piece. As such, a single element missing can lead to a wrong transcription, altering the composer’s musical intention. Therefore, it is very important to design transcription workflows to be robust and effective, to warrant a faithful digital recreation of a music composition. As such, the need for human evaluation and annotation becomes apparent, especially when we look closer at challenges that can arise from some real-life cases.

A first example of a challenging case is the use of different page structures for different orchestrations of a classical music piece. While transcriptions of a violin piece can have a system of measures with a single pentagram (see subsection 1.3.1 for details on these elements), a system of measures for a piano piece has two pentagrams to indicate notes for each hand. In the case of orchestral works, a system can cover a whole page with different pentagrams for each instrument used in the orchestration. This presents a fundamental challenge where the automatic recognition of the layout will need to be identified correctly, as it can impact the structure of the whole music score, passing wrong structural information to a potential XML representation.

A second (and perhaps the hardest) challenge for automatic recognition of music elements, is imposed by handwritten notations and scores. While modern printed or digital versions use specific lettering and spacing, which increase the readability of a music score, classical music composers have been composing by hand on paper for centuries. The page structure of those compositions can often vary, depending on how strictly a composer applied notational form. In some cases, the only challenge is the legibility of a composer’s inscriptions, while in others, we only have surviving sketches that hardly follow any standardized form (see Figure 1.3).



(a) Autograph score of Symphony no.95 by Joseph Haydn

(b) Autograph score of String Quartet Op.131 by Ludwig van Beethoven

Figure 1.3: Examples of handwritten music scores.

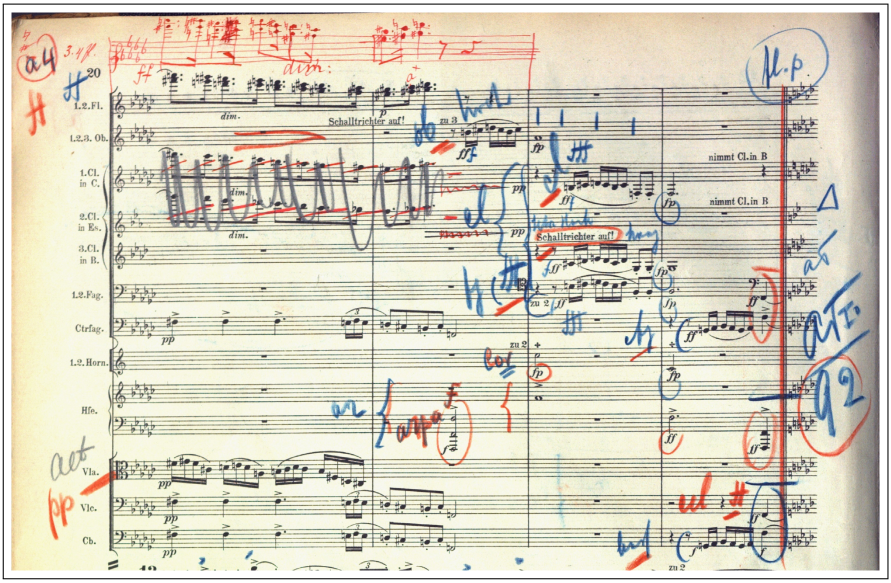


Figure 1.4: An example of handwritten notes overlayed on top of a printed score (Mahler 4th Symphony)

Similarly, we have surviving notes and alterations other musicians have made on top of existing scores. Those handwritten notes can vary from short descriptions to alterations of whole passages and scribbles of their interpretations (see Figure 1.4). These notes can be very important to preserve, as musicians can learn from others' musical perspectives, and music theorists can archive special interpretations of certain compositions.

1.4. An Opportunity for Hybrid Annotation Workflows

In this section, we showcase practices that can enable the inclusion of human annotation and assistance within algorithmic pipelines to overcome the challenges in the automatic processing of music compositions. These hybrid annotation workflows combine automatic and manual activities, where humans of different expertise can assist and enhance automated processes.

In a hybrid annotation workflow, the goal is to effectively and efficiently combine automated methods with human effort to achieve a result that couldn't be attained by either approach alone [55]. Individual workflow steps need to be identified early and, through careful design, allocated to the appropriate processing method. Automated methods and human input need to co-exist and complement each other. The complexity and (niche) knowledge required by a person to annotate and/or transcribe a music composition, but as the shortcomings of current automated methods, need to be factored in. Specifically, hybrid annotation workflows cover three main components: Algorithms / Machine Learning, the Crowd, and Quality Assessment. Figure 1.5 shows how these components interact to form end-to-end workflows.

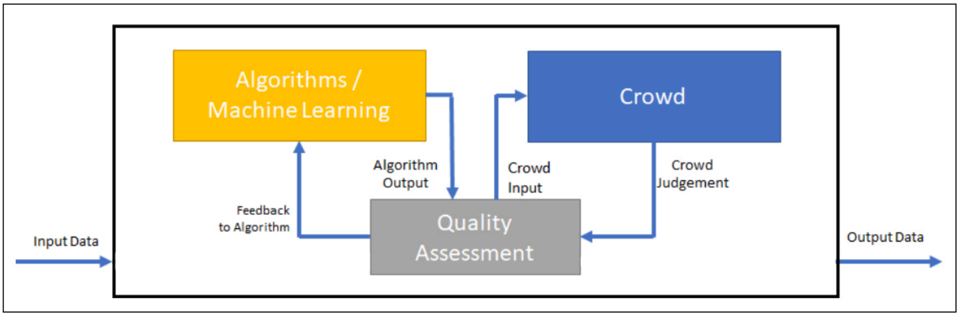


Figure 1.5: High-level Hybrid Processing Workflows

Algorithms / Machine Learning. A set of algorithms, typically machine learning algorithms, process the input data into the desired output data. These algorithms typically have at least one of the following two shortcomings:

1. The results produced by the algorithms are of bad quality and insufficient for the desired use.
2. The algorithm relies on extensive training data, which are typically not or only partially available.

The first problem can manifest in different ways [55]. For instance, the algorithm could be good enough in most cases but might fail in others. It would be necessary to identify when the algorithm fails (automatically or using crowdsourcing) and revert to using crowdsourcing fully redoing the failed task in these cases. The worst case would be a scenario in which the algorithm always fails, resulting in a pure crowdsourcing system. In another scenario, the algorithms generally work on all types of input data, but the output quality is slightly too low in nearly all cases. Here, all outputs need to be adjusted and fixed using crowdsourcing. The second problem requires creating training data that usually covers many examples of correct input data / desired output data pairs. Crowdworkers can provide such training pairs upfront for initial training or as part of the fixing measures introduced for the first problem. Then, this crowd-provided data can be used for incremental re-training.

Crowd. A crowd can be used to execute cognitive Human Intelligence Tasks. The choice of crowd workers and their incentivization is a core challenge, and while we do not address incentivization in this thesis, it could range from paying microtask workers on platforms like Amazon Mechanical Turk to motivating expert online communities using intrinsic incentives.

The profiling and modeling of annotators are very important components of any hybrid annotation system. By modeling annotators, we understand which of their attributes can affect their performance and how to assist them with purposeful and smart interface designs. We can also leverage task attributes to assign the "right" tasks to the "right" annotators whenever specific expertise and quality are needed. Therefore, in this thesis, we dedicate Chapter 2 to modeling and quantifying the crowd's musical affinity through several studies.

In general, a crowd can be utilized to:

1. Check the correctness of an intermediate algorithm result. This can range from simple correct/incorrect checks to more complex checks, which give a detailed overview of the location and nature of the error.
2. Produce results: Here, the crowd performs the same task the algorithm was designed for: transform a given input data instance into the correct output data. This functionality is employed when an algorithm fails to process, or when that data is required for further/initial training.
3. Improve results: Here, a machine-produced result with sub-par quality is manually improved. Typically, this should be employed when improving slightly faulty outputs is easier and cheaper than creating a new output manually from scratch.

Quality Assessment. This is a core component that ensures the effectiveness of a hybrid crowdsourcing process. The quality assessment is central in both judging the quality of algorithmic results and human annotations. Regarding assessing the algorithmic processes, the process can help decide if and what kind of crowd treatment is needed. On the other hand, regarding the human component, they can help judge the reliability and quality of crowd feedback in light of low-skill and/or malicious workers. Quality assessment strategies here could either mitigate the impact of such workers through sophisticated inter-annotator agreement schemes, attention tasks or through appropriate pre-screening and profiling of annotators.

1.5. CHALLENGES AND RESEARCH QUESTIONS

The main goal of this thesis is to gain insights into how to design solutions to scale up digitization and annotation methods of classical music compositions by including people of different expertise levels in different parts of the processes.

Throughout TROMPA we had the opportunity to collaborate with experts in classical music and music information retrieval. They played a valuable role during the requirement analysis for classical music transcription and provided input and resources that supported our efforts.

In the early stages of designing hybrid annotation workflows for classical music artifacts, we quickly became aware of the challenges of incorporating annotators of varying expertise. While experts' skills can be quantifiable through certificates, years of labour, relevant domain studies, and previous contributions, it is a hard challenge to quantify a non-expert's affinity for performing well in a music task.

Classical music composition can exist both in a visual and aural format, without one form necessitating the other (e.g., recordings of compositions with lost transcripts). Therefore, we needed a method that measures a person's capacity to "understand" music on multiple levels. Such a "music affinity" profile can help screen annotators for their capacity to perform a music transcription-related task and pair annotators to the right type of task appropriate to their capacity. Such capacity has to be quantified either by self-reporting questionnaires or by actively measuring it, to ensure its compatibility with other automated methods in a hybrid annotation workflow.

Based on these challenges, we were faced with the following research questions:

- **[RQ1]** How could the musical profile of annotators be characterized?
- **[RQ2]** How are different music perception skills and self-reported music-related knowledge distributed among non-experts?
- **[RQ3]** How are music perception skills associated with domain and demographic attributes?

While our design process of the musical competence framework is explained in Section 2.1 when tackling RQ2 and RQ3, we faced the challenge of identifying sources of non-experts that could readily exhibit musical affinity and potentially be available for annotating tasks. Crowdsourcing platforms, such as Prolific and Amazon Mechanical Turk, have proved reliable sources of available annotators so far, and as such, they were utilized for conducting our extensive user study in Section 2.2.

Borrowing from TROMPA's vision of rich, open, and democratized classical music repositories, we were also interested in identifying large communities outside crowdsourcing platforms that could also exhibit music affinity. A musical affinity that wouldn't be limited to only the popular compositions and composers of classical music, but would also expand our knowledge beyond the widely accepted "canon".

To identify such possibilities, we sought to tackle the following research question, conducting a community study on the Wikipedia and YouTube platforms in Section 2.3:

- **[RQ4]** How does the popularity of classical music composers on community-driven platforms relate to their album release trends?

While **Chapter 2** is dedicated to our work studying the musical characteristics of annotators and online communities, **Chapter 3** is presenting the challenges in designing and deploying annotation tasks for music transcription workflows, alongside our approach to address them.

More specifically, as part of our contributions to TROMPA we were presented with the engineering challenge of developing a system that would leverage the contributions of several annotators, to digitally recreate a semantically rich version of a scanned classical music score. As such, we had to face several difficult tasks, both regarding the transcription of the artifact itself but also the online availability of the system and its expected traffic. Section 1.3 gives an overview of music scores and challenges on automated transcription methods, while Section 3.1 showcases our workflow and design methodology behind our delivered system, Scriptoria. Limitations imposed by COVID-19 regulations hindered a wider engagement of users through TROMPA. Nevertheless, we could still live test Scriptoria online with amateur musicians (considered semi-experts in this thesis), contributing towards the transcription of the first pages of Ludwig van Beethoven's Sextet in E-flat major, op. 71.

Due to the use of semi-experts in the context of TROMPA, we were still met with open questions regarding the performance of non-experts in such extensive, end-to-end transcription workflows. Firstly, the symbols of music scores could prove discouraging to people not familiar with them. As such, we needed to experiment to what extent they could impose a challenge to non-experts and how could we possibly assist them.

Therefore, focusing on the task design for error detection (which could be part towards the end of a hybrid transcription workflow), in Section 3.2 we tackled the following research question:

- **[RQ5]** To what extent can workers from microtask crowdsourcing platforms detect errors in transcribed music scores?

Secondly, as discussed earlier in Section 1.1, a music score transcription can only be considered part of the network of information available for a given classical music composition. Different versions can coexist under the same composition entity, reflecting changes by the same composer or fellow musicians. Archiving and preserving those versions, though, can prove extremely challenging, especially if the transcripts are missing. This can be particularly true to versions of an original composition, where an instrumentalist or an orchestra made personal changes to a piece without publishing their version. In such cases, we are only left with audio or video recordings of their performances capturing their changes to the original composition. In the case of only audio recordings, we might even miss the information of the exact instrument used for the performance, a piece of crucial information as most classical music compositions were written with a specific instrumentation in mind.

An expert or trained musician is expected to relatively easily recognise the notes or instruments involved in an audio recording. However, little research has investigated the ability of untrained individuals to identify these elements in an audio recording.

While single-instrument excerpts have been tested previously on crowdsourcing platforms [37], we have little understanding of instrument recognition on polyphonic excerpts. With this challenge in mind and looking beyond the classical music genre towards genres with similar challenges, we finally studied the following research questions in Section 3.3, by designing and testing appropriate task interfaces:

- **[RQ6]** To what extent can non-experts detect the onset and offset of a musical instrument's activity on polyphonic audio?
- **[RQ7]** How do their self-assessed perceptual abilities and musical knowledge relate to their performance?

1.6. Contributions and Origins of Chapters

This thesis is thematically organized into four sections. They present: (a) the motivation and context behind the research questions; (b) the studies conducted to understand and model music annotators; (c) the solutions we designed and employed to tackle challenges in processing classical music compositions; and finally (d), how our findings expand our understanding of knowledge-intensive tasks and how they can be applied beyond classical music.

The research and studies conducted throughout this thesis, have been published in peer-reviewed scientific venues, relevant to either music information, web engineering or human computation. The manuscript also includes relevant theoretical and practical work that was partaken by the author throughout the duration of the European H2020 project TROMPA, reviewed and approved by project partners and the appointed European committee.

In detail:

In **Chapter 1**, Section 1.4 and parts of Section 1.3 are based on content from a scientific paper published in the 3rd International Workshop on Reading Music Systems (WoRMS) [83]. This chapter acts as an introduction, contextualizing the work in this thesis, providing background information on classical music transcription and recognition methods, while also contributing a high-level design of Hybrid Annotation Workflows.

In **Chapter 2**, Section 2.1 is based on relevant excerpts from work conducted and published on deliverables for the European H2020 project **T**owards **R**icher **O**nline **M**usic **P**ublic-domain **A**rchives (TROMPA) while Section 2.2 is based on a full research paper published in Human Computation and Crowdsourcing (HCOMP) conference [86] and a journal paper published in Frontiers in Artificial Intelligence [87]. Finally, Section 2.3 is based on a research paper published in the International Conference on Web Engineering (ICWE) [85]. This chapter contributes a theoretical framework to model musical properties of annotators; an extensive study on musical sophistication and perceptual capabilities of crowdworkers; and a study on community-driven platforms to understand user engagement with classical music composers.

In **Chapter 3**, Section 3.1 is based on a demonstration paper published in The Web Conference (TheWebConf) [82], expanded with detailed insights on the process from our contributions to TROMPA. Sections 3.2 and 3.3 are based on two full research papers published in International Society for Music Information Retrieval (ISMIR) [88, 84]. The chapter contributes to the design of a deployed, online system of microtask-based music score transcription, alongside two studies on task-specific crowdsourcing experiments, music score error detection and temporal activity detection of music instruments in audio.

The final chapter (**Chapter 4**) summarizes the main findings in this thesis while presenting possible future directions that can help to further improve human computation workflows for knowledge-intensive music annotation tasks.

2

MODELING AND MEASURING THE MUSICAL AFFINITY OF CROWD ANNOTATORS

How can we quantitatively capture a person's musical competence? Are self-reported, music perception skills reliable? Do people online generate valuable information regarding classical music through engagement? These are some of the questions that inspired our research in this chapter. We present our work in modeling musical competence, to be used, for instance, to assign transcription tasks to annotators. We also present our extensive study in identifying non-experts online that can readily exhibit musical affinity, alongside our exploration of how classical music information is generated in community-driven platforms. Results and insights of these works provide a novel understanding on the music affinity of annotators with varying levels of expertise, enabling their reliable inclusion in demanding music transcription workflows.

This chapter is based on excerpts published in deliverables during the European H2020 project ***T****owards* ***R****icher* ***O****nline* ***M****usic* ***P****ublic-domain* ***A****rchives, and the following publications: "Exploring the music perception skills of crowd workers" [86], "An Analysis of Music Perception Skills on Crowdsourcing Platforms" [87] and "On the Popularity of Classical Music Composers on Community-Driven Platforms" [85].*

2.1. Musical Properties of Annotators

Due to the diversity of the use case scenarios defined during TROMPA, it was expected that there will be a large variety of users/contributors with diverse backgrounds, competences, and experience levels in the domain of music. Similarly, it was also expected that the types of content annotation tasks that would be distributed to those users, would vary significantly and require different levels of knowledge and skills; some expecting general knowledge on a topic, while others needing more niche specializations. To model this, we adopted and expanded a specialized competence model introduced in Juri, L., et al, 2007 [42], to tackle our first research question:

- **[RQ1]** How could the musical profile of annotators be characterized?

The core goal of our modified competence model was to enable appropriate assignments of users to tasks based on their competences and the needs of a given task. Therefore, our music competence model would have to model the user's musical abilities in a way that we can computationally "connect" them to specific types of tasks. To that end, we defined "Profiles" with a set of competences, which are at the core of the Music Competence model. The versatility of our model lies in the fact that "Profiles" can be used to both describe the user's competences, but also the competences required to perform a task. Using the same class profile for both users and tasks allows for abstracting many computational tasks related to competences, such as computing a competence gap. Competence gap analysis compares two sets of profiles (the target and the available) to determine their difference. The same implementation of this functionality then can be used for many different purposes, e.g., to check whether a user has the required competences to perform a task, to compare the skills of two users, to compare the complexity of skills, or to find the best task for each user. This model was designed and delivered to enable TROMPA partners to design a task, create a "Profile" of competencies needed to perform it, and then the system can rely on competence gap analysis to find the best users for the task. Specific skills associated with specific competences, would be determined by expert project partners in the music field.

Finally, based on our design, each user can have multiple competence profiles, which are obtained from a variety of sources (as we expand in sub sections 2.1.1 and 2.1.2). Each of these profiles could be time-stamped, while the actual competences a user possesses, could be computed by unifying all profiles and resolving conflicts (i.e., competences referring to the same skill but different proficiencies or contexts) based on the timestamp and the sources of the profiles. While the final infrastructure of TROMPA didn't facilitate the online deployment of the music competence model, our designs were used to communicate task design concepts and annotator properties within the consortium, but also to inspire our user studies presented in this thesis.

2.1.1. Modelling Musical Competence

For the purposes of musical competences, we adopt the definition of competence as "effective performance within a context at different levels of proficiency", as given in Cheetham, G. and Chivers, G.E., 2005 [19]. We see competences as tuples of "Skill" (i.e., the ability to perform a certain activity) and "Proficiency" (i.e., a measure of how well

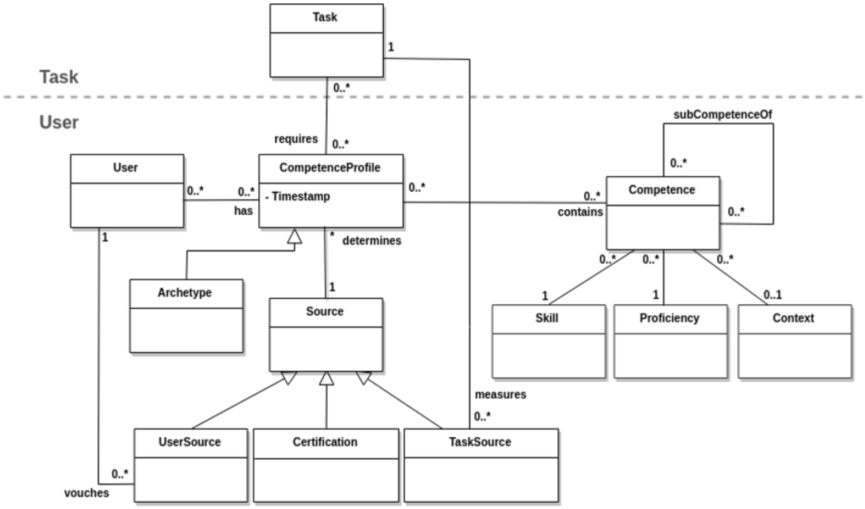


Figure 2.1: Modeling annotator competence and task requirements

the skill can be performed), while optionally attaching "Context" in which the skill and proficiency are relevant. For example, "Professional classical piano" would be composed of the skill "Piano playing", the proficiency level "Professional" and the context "Classical music". In addition, we also allow for modeling sub-competences for adding granularity based on the needs of a given task. Below we expand on the concepts of our Music Competence model, while a graphical overview is given in Figure 2.1 to accompany our descriptions.

Skill: Skills describe reusable domain knowledge, or the ability to perform a certain activity. This could be a complex skill like "Piano playing", but also a simpler skill like "Reading musical score sheets". If needed, sub-competences can be modelled, i.e., a complex competence can be composed of simpler ones. The Competence Model is compatible with skill taxonomies, enabling domain experts to define such taxonomies which fit their content annotation requirements.

Proficiency: Proficiency describes the quality or extent to which a user possesses a skill. Many different scales may be used, but it should be possible to reuse them within and across the borders of a use case scenario. For example, scales are typically the same for most certifications given by similar institutions (e.g., all curricula within a country's universities). Hence, they can be modeled once and referenced many times. Competence descriptions can refer to specific items of these scales in order to represent the proficiency level acquired/required. Algorithms could take relationships among proficiency levels into account in order to assess how much training/learning is required to reach a determined proficiency level.

Context: Context is commonly defined as "the interrelated conditions in which something exists or occurs". Regarding competences, context may refer to different concepts. It might be the specific occupation in which a competence is acquired (e.g., playing piano in a conservatorium), a set of topics within a domain (e.g., classical music or folk music) or even the personal settings related to the learner (e.g., competencies are different if acquired in a group-based learning setting than individually). Modeling contexts may be a complex task, as it may coincide with modeling the whole domain of knowledge of an institution. So far, our investigations of existing relationships between context elements (regarding its use within competences) do not show the need for providing such complex representation. For this reason, we decided to keep the implementation simple but easily extensible.

Archetypes: Archetypes are a convenience class introduced in the model to support the process of adding new users and setting up their initial competence profiles based on predefined groups of expected competences. As such, an archetype describes a set of competences which are frequently acquired together. An example archetype could be an MSc degree in a given course of studies: while each individual student might have slightly different skills with slightly varying proficiencies based on their personal curriculum choices and performance, a given master's degree typically represents a selected set of competences. Another example of an archetype could be occupational or job-related archetypes, like "being a professional symphonic orchestra conductor", which would imply a large variety of different skills like good baton techniques and/or advanced aural skills, which are typically shared between all people in this occupation. Furthermore, we can also model archetypes to represent the implicit competences expected from members of different communities, thus linking the competence to a social model. Thus, by modelling an extensive body of archetypes for TROMPA, a new user's profile could be quickly setup by importing and copying all archetype profiles which apply to them. Their profile could then be further personalised by expanding their inherited competences, with additional competences that describe individual deviations of that user from the more generic archetype profiles initially assigned to them. As such, the archetypes will be helpful in defining large user groups, as several users might share the same set of competences, while keeping the model's versatility by accommodating individual sets of skills and proficiencies.

2.1.2. SOURCES AND VERSIONING

Each user can have several profiles attached to them. As previously stated, when a new user joins the system, we envision that they select all archetypes which apply to them (like archetypes referring to formal education or those related to a specific profession or occupation) and may also self-report additional specific competences. To keep track of the lineage of each competence profile, profiles are time stamped and can also have an optional "Source" attached to them. The "Source" describes how the profile was obtained, and/or what vouches for the validity of the the profile. These profile sources can be used for reliability assessment of the data obtained from annotators when evaluating their profiles.

In the current model, we include three types of sources: "Users", "Tasks", and "Certifications". A "User" source describes that a user vouches for their competence profile (when self-reporting competence profiles) or the competence profile of another user (i.e., in a scenario where users know each other or come in contact with each other within their social community). A "Certification" source refers to a profile which was obtained by having a formal certification of any type, typically issued by an external organization. This can be used for example for archetype profiles representing degrees in formal education by attaching additional information on type, time, and organization. The last source for competence profiles are "Task"-sourced profiles which are created by observing users while conducting music annotation tasks. For example, if for a task requiring a certain competence a user performs particularly well (or even beyond the expectations based on their current competence profile), a new profile with updated competence proficiency levels can be created to reflect this. The same applies when a user should have certain competences, but repeatedly fails to perform tasks with matching competences adequately. These new "Competence Profiles" are timestamped to allow analysis of data lineage, when assessing the historical evolution of the community of contributors.

2.2. MUSIC SOPHISTICATION AND AURAL PERCEPTION SKILLS OF THE CROWD

Several studies have shown the ability of crowd workers to successfully contribute to the analysis and annotation of multimedia content, both based on simple perceptual skill, e.g., for image analysis [97] and domain-specific knowledge, [70]. Musical content is no exception, and research has shown that the general crowd can be successfully involved in the annotation [88] and evaluation [104] processes of music-related data and methods. Plenty of music annotation tasks, [49, 51, 59, 50, 98] can be routinely found on microtask crowdsourcing platforms, mostly focused on descriptive [47] and emotional [49] tagging.

Music, as a form of art, often requires a multifaceted set of skills to perform and certain expertise to analyse its artifacts. There are cases that require advanced music perceptual skills (such as the ability to perceive changes in melody) and music-specific knowledge. However, both in literature and in practice, it is rare to encounter such crowdsourcing tasks. Consider, for example, annotation tasks targeting classical music, e.g. music transcription, performance evaluation, or performance annotation. Classical music is a genre featuring artworks with high musical complexity; it is no surprise that corresponding analysis and annotation tasks are often exclusively performed by musical experts and scholars. This unfortunately hampers current efforts to digitize and open up classical music archives, as scholars and experts are expensive and not easily available. Here, the ability to utilise microtask crowdsourcing as an annotation and analysis approach could bring obvious advantages. But how likely it is to find advanced music-related perceptual skills on crowdsourcing platforms? With the goal of answering this broad research question, in this paper we scope our investigation on the following two aspects:

- **[RQ2]** How are different music perception skills and self-reported music-related knowledge distributed among crowd workers of different platforms?
- **[RQ3]** How are music perception skills associated with domain and demographic attributes?

Studies on human cognition and psychology have shown that people can possess innate music perception skills without previous formal training [60, 103]. However, the majority of those studies have been conducted in labs, under controlled conditions and with limited amounts of participants.

In our work we set out to measure the music sophistication and perception skills of crowd workers operating on the Prolific[1] and Amazon Mechanical Turk[2] crowdsourcing platforms. We chose to conduct our study on these two different platforms, in order to diversify our participant pool and identify potential differences between them. In its present form, this study expands the preliminary study as presented in [86], by diversifying the participant pool and complementing the analysis with additional methods.

We designed a rigorous study that employs validated tools to measure the musical sophistication of the users and quantify their music perception skills: the Goldsmith's Music Sophistication Index (GMSI) questionnaire [65] and the Profile of Music Perception Skills (PROMS) active skill test [48] respectively (and more specifically its shorten version: Mini-PROMS). These tools allow for a general overview of musical ability characteristics, but also a more detailed understanding through their subcategories (e.g. musical training and melody perception skills). By juxtaposing passive methods of assessment (questionnaire) with the active evaluation of auditory skills, we aim to gather a better understanding of workers' actual skills on musical aspects, beyond their subjective self-assessment. With GMSI, we are able to evaluate a person's ability to engage with music through a series of questions focusing on different musical aspects. PROMS on the other hand, allows for a more objective way to measure a person's auditory music perception skills (e.g. melody, tuning, accent and tempo perception) through a series of audio comparison tests. To the best of our knowledge, this is the first attempt to use PROMS in an online crowdsourcing environment and the measured perception skills can offer valuable insights to the auditory capabilities of the crowd.

Our findings indicate that pre-existing musical training is not common among crowd workers, and that music sophistication aspects are not necessarily predictive of actual music perception skills. Instead, we observe that the majority of workers show an affinity with specific sets of skills (e.g., we found a surprising number of *musical sleepers* — workers without formal training but still high music perception skill test results). As a whole, our study paves the way for further work in worker modelling and task assignment, to allow a wider and more refined set of microtask crowdsourcing tasks in the domain of music analysis and annotation.

2.2.1. Related Work

There is a long history of studies on perception and processing of music by humans; from the analysis of the socio-cultural variables influencing a person's amplitude for musicality [34], to the study of musicality from a genetics' base [29]. In all cases, inherent music processing capabilities have been found in people and they seem to be connected with basic cognitive and neural processes of language since early stages of development [52, 45]. Even people with *amusia*, a rare phenomenon where a person can't distinguish tonal differences between sounds [74], they can still process and replicate rhythm correctly [38].

In [65], we find a large scale study on musical sophistication through the use of the GMSI survey, on a unique sample of 147,663 people. GMSI is particularly calibrated to identify musicality in adults with varying levels of formal training. It is targeted towards

[1] https://www.prolific.co

[2] https://https://www.mturk.com

the general public, and can prove less effective to distinguish fine differences between highly trained individuals. Musical sophistication in the context of that study, and ours, encompasses musical behaviours and practices that go beyond formal training on music theory and instrument performance. Their findings show that musical sophistication, melody memory and musical beat perception are related. The survey has been translated and replicated successfully (on smaller samples) in French[21], Portuguese[53], Mandarine[54], and German[90].

Our study draws connections to those findings and aims to shed light into the musical capabilities of people on crowdsourcing platforms. The demographics and conditions of the studies presented so far, cannot be easily compared to those of online markets. Users on those platforms are participating in such studies through monetary incentives, and the conditions (equipment, location, potential distractions, etc.) under which they perform the tasks cannot be controlled as in a lab environment [102, 112, 26].

Currently, crowdsourced music annotation is primarily utilised for descriptive [47] and emotional [49] tagging. Large-scale music data creation and annotation projects such as Last.fm[3] and Musicbrainz[4], are largely depended on human annotation, but from users of their respective online social platforms. A survey on the applicability of music perception experiments on Amazon Mechanical Turk [68], showed that online crowdsourcing platforms have been underused in the music domain and the status has not changed radically since then. Through our study, we want to examine the capabilities of the crowd on processing music audio and showcase their capabilities, in an attempt to encourage further research and utilisation of crowdsourcing in the music domain.

2.2.2. Experimental Design

The main focus of this study is to offer insights into the musical characteristics and perception skills of workers operating on crowdsourcing platforms. We therefore designed our experiment to capture these attributes through methods that can be used online, and that do not require pre-existing musical knowledge. We used two methods: 1) the *GMSI* questionnaire to evaluate the *musical sophistication* (musical training, active engagement and other related musical characteristics) [65]) of workers and 2) the *Mini-PROMS* test battery to evaluate their auditory music perception skills. We then compare the obtained results, paying specific attention to the overlapping aspects of musical sophistication and music perception skills. With this experiment, we are also interested in identifying "*musical sleepers*" and "*sleeping musicians*", a notion originally presented in [48]. A musical sleeper is a person with little to no musical training but with high performance in the perception test, while a sleeping musician indicates the opposite.

Procedure

After a preliminary step where workers are asked basic demographic information (age, education, and occupation), the study is composed of four consecutive steps (Figure 2.2), each devoted to collecting information about specific attributes corresponding to the crowd workers: 1) Musical Sophistication Assessment (*GSMI*), 2) Active Music Perception

[3] https://www.last.fm

[4] https://musicbrainz.org

Skill Assessment (*Mini-PROMS*) and 3) Post-task Survey collecting information on workers audio-related conditions, and perceived cognitive load.

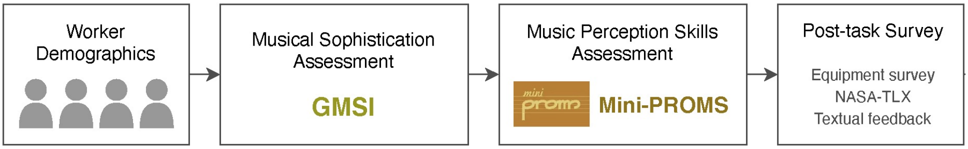


Figure 2.2: The four steps in the music perception skills study.

QUESTIONNAIRES AND MEASURES

Capturing Musical Sophistication of Workers. Musical behaviours of people such as listening to music, practicing an instrument, singing or investing on vinyl collections, all show the affinity of a person towards music. The degree to which a person is engaged to music through these behaviours, constitutes the musical sophistication. Musical sophistication can be measured as a psychometric construct through the *GMSI* questionnaire, which collects self-reported musicality through emotional responses, engagement with music, formal training, singing capabilities and self-assessed perception skills. It is an instrument specifically designed to capture the sophistication of musical behaviours, in contrast to other questionnaires such as Musical Engagement Questionnaire (MEQ) [108], which measures the spectrum of psychological facets of musical experiences. More specifically, the musical sophistication of people based on [65], is organised into the following five facets:

- Active Engagement: this aspect determines the degree to which a person engages with music, by listening to and allocating their time/budget to it;
- Perceptual Abilities: this aspect assesses the skill of perceiving (mainly auditory) elements of music. This is an important subscale in our study, since the self-assessed perceptual skills of the workers in GMSI can be directly compared to those we actively measure in Mini-PROMS;
- Musical Training: this aspect reports the years of training on aspects of music (e.g. theory, performing an instrument), which can indicate the formal expertise that a person has in the domain;
- Emotions: this aspect determines the emotional impact of music on that person;
- Singing Abilities: this aspect evaluates the ability to follow along melodies and tempo (beat) of songs.

GMSI offers additional questions outside the subscales, which capture specific properties of the participant: 1) “Best Instrument”, which represents which instrument the user knows to play the best, 2) “Start Age”, which age the participant starting learning an instrument and 3) “Absolute Pitch”, which indicates if the person can understand correctly the exact notes of a sound frequency. Absolute pitch is a very rare trait that develops during the early stages of auditory processing [15] but can deteriorate through the years [5]. As such, a person with perfect pitch perception, could have an advantage on a melody perception test, thus we included it with the rest of the subscales.

The original GMSI questionnaire contains 38 main items and 3 special questions, and considering the rest of the study's parts, we chose to reduce its size while keeping its psychometric reliability. For that purpose, we consulted the GSMI online "configurator"[5] which allows to select the number of items per subscales and estimates the reliability of the resulting questionnaire based on the questions it selects. We reduced the size of the questionnaire to 34 questions, and preserved the special question about "Absolute Pitch", resulting in 35 questions in total.

In the GMSI questionnaire each question from the subscales uses the seven-point Likert scale [41] for the user's responses, with most questions having "Completely Agree", "Strongly Agree", "Agree", "Neither Agree Nor Disagree", "Disagree", "Strongly Disagree" and "Completely Disagree" as options. Few questions offer numerical options for topics (e.g. indicating the time spent actively listening to music, or practicing an instrument). The workers are not aware of the subscale each question belongs to. The index of each subscale of GMSI is calculated with the aggregated results of the relevant questions. The overall index of "General Music Sophistication" is calculated based on 18 questions out of the total 34 items of the subscales; these 18 questions are predefined by the designers of the questionnaire; the question about "Absolute Pitch" does not contribute to the total index.

Using the GMSI questionnaire is close to the typical methods used to assess the knowledge background of annotators in other domains. Especially the questions of "Musical Training" follow standard patterns to assess the formal training of a person in a domain, thus a certain objectivity can be expected (assuming good faith from the workers). However, the rest of the categories are based purely on subjective indicators and self-reported competence, which can potentially misrepresent the true music behaviours and capabilities of a worker. For this reason, it is necessary to understand the best practices that could reliably predict a worker's performance to a music annotation task. To that end, we compare the workers' input in such questionnaires, and specifically on GMSI, to the music perceptual skills they might possess, which we measure through an audio-based, music perception skill-test.

Measuring Music Perception Skills of Workers. The music perception skill test is based on the well-establish *Profile of Music Perception Skills* (PROMS) test, [48]. Its original version is quite extensive and its completion can take more than an hour, as it covers several music cognition aspects like Loudness, Standard rhythm, Rhythm-to-melody, Timbre, Pitch and more. Considering the possibly low familiarity of crowd workers with these tasks and its inherent difficulty, we opted for a shorter version, the *Mini-PROMS* [111], which has also been adopted and validate in the context of online, uncontrolled studies.

Mini-PROMS is a much shorter battery of tests (15 minutes completion time), which still covers the "Sequential" and "Sensory" subtests. It can measure a person's music perception skills, by testing their capability to indicate differences on the following musical features:

[5] https://shiny.gold-msi.org/gmsiconfigurator/

- Melody: A sequence of notes, with varying density and atonality
- Accent: The emphasis of certain notes in a rhythmic pattern
- Tuning: The certain frequency of notes, when played in a chord
- Tempo: The speed of a rhythmic pattern

The musical aspects selected in this test are argued to well represent the overall music perception skills of a person, only in a more concise way. This version retains test–retest reliability and internal consistency values close to the original PROMS test [48], validating it for our research purposes. Note that, although reduced in size, these four skills are required to enable a broad range of music-related research, such as beat tracking, tonal description, performance assessment and more.

For each of the 4 musical aspects workers receive a brief explanation and an example case to familiarise the user with the test. Each test after the introduction, presents a reference audio sample twice and a comparison sample once. The two audio samples can differ based on the musical aspect tested and the worker is asked if the samples are indeed same or differ. The authors of PROMS have put particular effort on distinguishing the musical aspects from each other, to make the skill evaluation as close as possible to the musical aspect tested. Finally, to minimise cognitive biases due to enculturation, [22], the audio samples have been created using less popular instrument sounds, such as harpsichord and "rim shots". Meanwhile, the structure of audio samples and the aspect separation allow for a more precise measurement of a person's perception skill.

The categories of “Melody” and “Accent” have 10 comparisons each, while “Tuning” and “Tempo” have 8. After the user has listened to the audio samples, they are asked to select between “Definitely Same”, “Probably Same”, “I don't know”, “Probably Different” and “Definitely Different”. The participant is then rewarded with 1 point for the high-confidence correct answer, while the low-confidence one rewards 0.5 point. The subscale scores are calculated through a sum of all items within the scale and divided by 2. The total score is an aggregated result of all subscale scores. During the test, the user is fully aware of the subscale they are tested for, but the name of “Tempo” is presented as “Speed” (original creators' design choice).

Self-assessment on Music Perception Skills. Self-assessment can often misrepresent an individual's real abilities[46]. For that reason, we employed a survey to study this effect its manifestation with music-related skills. After Mini-PROMS test, the worker has to input how many of the comparisons per subscale they believe they correctly completed - this information is not known to them after executing the Mini-PROMS test. Therefore, they are presented with 4 questions, where they have to indicate between 0 and the total number of tests per subscale (10 for “Melody”/“Accent” and 8 for “Tuning”/“Tempo”). Finally, the results of this survey, are compared to the score of workers on the “Perceptual Abilities” subscale of GMSI, which also relies on self-assessment. We expect workers to re-evaluate their own skills, once exposed to the perception skill test.

Post-task Survey. As a final step of the task, the worker is presented with three post-task surveys: 1) a survey on the audio equipment and the noise levels around them, 2) a survey on the cognitive load they perceived and 3) an open-ended feedback form.

The audio equipment survey consisted of four main questions, to retrieve the type of equipment, its condition and the levels of noise around them during the audio tests. Insights on these can help us understand to what extent the equipment/noise conditions affected Mini-Proms test, which is audio-based. More specifically, we asked the following questions:

1. What audio equipment were you using during the music skill test?
2. What was the condition of your audio equipment?
3. Does your audio equipment have any impairment?
4. How noisy was the environment around you?

The options regarding the audio equipment were: “Headphones”, “Earphones”, “Laptop Speakers” and “Dedicated Speakers”. For the condition questions (2) and (3), we used the unipolar discrete five-grade scales introduced in [77], to subjectively assess the sound quality of the participants’ equipment. Finally, for question (4) on noise levels, we used the loudness subjective rating scale, introduced in [9].

In the second part of post-task survey, the workers had to indicate their cognitive task load, through the NASA’s Task Load IndeX (NASA-TLX) survey[6]. The survey contains six dimensions — Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Workers use a slider (ranging from 0 to 20, and later scaled to 0 to 100) to report their feelings for each of the six dimensions. A low TLX score represents the music skill test is not mentally, physically, and temporally demanding, and it also indicates less effort, and less frustration perceived by the worker, while completing the entire study.

Finally, we introduced a free-form textual feedback page, where users were encouraged to leave any comments, remarks, or suggestions for our study.

Worker Interface

The worker interfaces of our study is using VueJS[7], a JavaScript framework. The first page of our study, contained general instructions for the study alongside estimated completion times for each part of it. Each page thereafter, contained an interface for each of the steps in our study, as seen in Figure 2.2.

To assist navigation through the GMSI questionnaire, we implemented the questionnaire interface to show one question at a time. We added a small drifting animation to show the next question, when they select their answer in the previous one. We also added a “back” button, in case they wanted to return to a previous question and alter their answer. They could track their progress through the questionnaire from an indication of the number of the question and the total number of questions (see Figure 2.3).

While we retrieved the questions for GMSI and implemented them in our study’s codebase, for PROMS we wanted to use the exact conditions and audio-samples as in [111]. To replicate their test faithfully, the creators of PROMS [48] kindly gave us access to their Mini-PROMS interfaces (example interface in Figure 2.3). Mini-PROMS

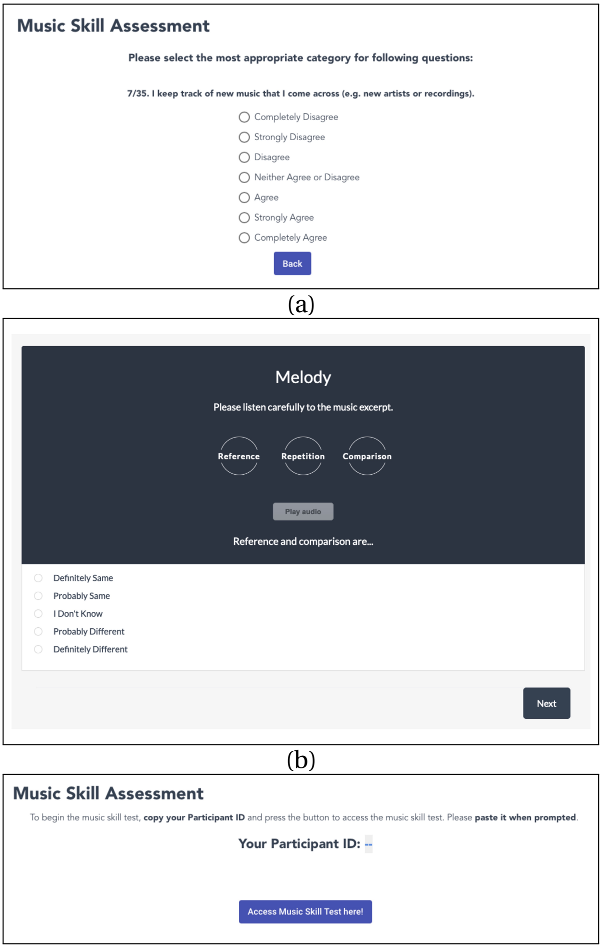
[6] `https://humansystems.arc.nasa.gov/groups/tlx/`

[7] https://vuejs.org

is implemented on LimeSurvey[8] and users were redirected to it after the completion of GMSI.

After the GMSI questionnaire, workers were introduced to the page seen in Figure 2.3. There, they had to copy their Participant ID (retrieved programmatically from the crowdsourcing platforms) and use it in the Mini-PROMS interface later, so we could link their test performance (stored in LimeSurvey), with their entries in our database. At the end of Mini-PROMS, the users were redirected back to our study through a provided URL.

In the final stage of our study, the participants were greeted and provided a "completion code", which they could submit on back on their respective platform, to complete the task.



Music Skill Assessment

Please select the most appropriate category for following questions:

7/35. I keep track of new music that I come across (e.g. new artists or recordings).

○ Completely Disagree
○ Strongly Disagree
○ Disagree
○ Neither Agree or Disagree
○ Agree
○ Strongly Agree
○ Completely Agree

(a)

Melody

Please listen carefully to the music excerpt.

Reference and comparison are...

○ Definitely Same
○ Probably Same
○ I Don't Know
○ Probably Different
○ Definitely Different

(b)

Music Skill Assessment

To begin the music skill test, **copy your Participant ID** and press the button to access the music skill test. Please **paste it when prompted**.

Your Participant ID: --

(c)

Figure 2.3: Interfaces of the study (a, GMSI questionnaire, b, Mini-PROMS, c, Participant ID prompt.

[8] https://www.limesurvey.org

Participants, Quality Control, and Rewards

On Prolific, we recruited 100 crowd workers to complete our study. We applied a participant selection rule for "Language Fluency": English, as all of our interfaces were implemented in English. Only crowd workers whose overall approval rates were higher than 90% could preview and perform our study. On Amazong Mechanical Turk, we recruited 100 crowd workers as well, where we set their approval rate to "greater than 90%".

To assess the quality of the user input, we included attention check questions on the GMSI and NASA-TLX interfaces of the study. More specifically, we included three attention check questions in GMSI, asking the participants to select a specific item in the same seven-point Likert scale. In the NASA-TLX survey, we included a question asking the users to select a specific value out of the 21 available in the scale of the survey.

We set the reward on Prolific and Amazon Mechanical Turk for completing our study to 3.75 GBP (5.2 USD). Upon the completion of our study on both platforms, workers immediately received the reward. The average execution time was 32.5 minutes, resulting in the hourly wage of 7.5 GBP (10.3 USD), rated as a "good" pay by the platforms.

2.2.3. Results

While investigating the data we gathered in our study, we followed similar analysis steps for both platforms. The data were first cleaned up based on our attention check questions and we only kept demographic data that we had actively asked the participants (dropped platform-based demographics).

We proceeded with identifying the distribution characteristics of each variable from the different parts of our study (GMSI and Mini-PROMS subcategories, NASA-TLX and equipment questions). Combined with the intercorrelations per study part, we gained important insights on the attributes of each variable and their relations. These results are compared to those of the original GMSI and Mini-PROMS studies, to assess the differences between the different participants' pools. Finally, we run a Multiple Linear Regression, to assess which factors seem to be the best predictors for the music perception skills of a crowd worker (e.g. musical training, equipment quality etc.).

Prolific

Of the 100 workers recruited from Prolific, 8 of them failed at least one attention check question(s); 5 of them provided invalid/none inputs. After excluding these 13 invalid submissions, we have 87 valid submissions from 87 unique workers.

Worker Demographics Table 2.1 summarises workers' demographic information. Of the 87 crowd workers who provided valid submissions, 36 were female (41.38%), while 51 were male (58.62%). Age of participants ranged between 18 and 58 and the majority of them were younger than 35 (87.36%). The majority of the workers (51%) were reported to be unemployed, while from those employed, 73.17% had a full-time job. Most workers had enrolled for or acquired a degree (78.16%), with 51.47% of them pointing to Bachelor's degree. In total, we employed workers from 15 countries, with most workers (77%) currently residing in Portugal (25), United Kingdom (16), Poland (13) and South Africa (13).

	Variables	**Statistics**
Age (years)	Range	18-65
	Majority	18-25 (70.11%)
Occupation	Full-time	30
	Part-time	11
	Unemployed	44
	Voluntary Work	2
Education	Associate degree	3
	Bachelor's degree	35
	Doctorate degree	1
	High school/HED	16
	Master's degree	12
	Professional degree	1
	Some college, no diploma	13
	Some high school, no diploma	2
	Technical/trade/vocational training	4

Table 2.1: Prolific participant demographics

Results on Worker Music Sophistication Table 2.2 summarises the results of the GMSI questionnaire on our workers. We contrast our results to results of the original GMSI study [65], which covered a large population sample of participants $n = 147,663$ that voluntary completed the questionnaire, on BBC's *How Musical Are You?* online test. Participants were mainly UK residents (66.9%) and, in general, from English-speaking countries (USA: 14.2%, Canada: 2.3%, Australia: 1.1%), with 15.9% having non-white background. The sample contained a large spread on education and occupation demographics, where only 1.8% claimed working in the music domain. To some extent, this study is considered representative for the general population in the UK (but is biased towards higher musicality due to the voluntary nature of that study). As such, we can assume a certain disposition and affinity to music from GMSI's population sample, compared to ours where the incentives where monetary.

In our study, the observed General Music Sophistication ($\mu = 69.76$) positions our workers pool at the bottom 28-29% of the general population distribution found in the GMSI study. We observe a similar effect also with the individual subscales with the exception of "Emotions", for which our workers fare a bit higher (bottom 32-38%).

The result indicates that the self-reported music sophistication of crowd workers is strongly below that of the general population. Most workers had received relatively little formal training in their lifetime. This finding is important for the rest of the analysis, as it indicates *low formal expertise* with music among the crowd workers.

Most workers indicate relatively high perceptual abilities ($\mu = 33.62$, $max = 45$). Here, it is interesting that previous studies [6] estimate that less than 1% (or 5 people) per 11,000 possess "Absolute Pitch". In our sample though, 9 workers indicated having this characteristic, little more than the 10% of our sample. This could indicate a possible confusion between quasi-absolute pitch which is related to the familiarity of a person with

	Range	Median	Mean	Standard Deviation(1σ)
Active Engagement	19-45	31	30.91	5.45
Perceptual Abilities	16-45	34	33.62	6.65
Musical Training	7-45	17	18.52	9.61
Singing Abilities	9-41	28	27.41	6.03
Emotions	18-42	33	33.24	4.28
General Music Sophistication	40-101	69	69.76	14.20

Table 2.2: GMSI Range, Median, Mean and Standard Deviation

an instrument's tuning and timber [78], or with relative pitch. Relative pitch is trainable through practice and useful to professional musicians, as they can detect changes in pitch through the relations of tones (5 out of 9 workers who indicated "Absolute Pitch" had scored higher than 30 out of 49 in the "Musical Training" category scale, indicating adequate formal musical training).

Table 2.3 presents the correlations between GMSI subscales. As the scores of each GMSI subscale follow a normal distribution (Shapiro-Wilk test), we applied Pearson's R test to calculate correlation coefficients. We observe that Perceptual Abilities shows positive correlations with most other subscales ($p < 0.05$), especially with Music Training ($R = 0.442$), Emotions ($R = 0.380$), and Singing Abilities ($R = 0.463$). This finding suggests that the listening skill plays the most important role in crowd workers' music sophistication. We also find significant correlations between Active Engagement and Emotions ($R = 0.401$), and between Singing Abilities and Musical Training ($R = 0.465$). The original GMSI study has shown that different subscales are strongly correlated ($R > 0.486$). The difference we observe could be partly explained by the generally lower musical sophistication scores of the crowd workers in our pool.

	Active Engagement	Perceptual Abilities	Musical Training	Emotions	Singing Abilities
Active Engagement	1.000				
Perceptual Abilities	0.262*	1.000			
Musical Training	0.224*	0.442*	1.000		
Emotions	0.401*	0.380*	0.178	1.000	
Singing Abilities	0.142	0.463*	0.465*	0.125	1.000

Statistical significance ($p < 0.05$) is marked using an asterisk (*).

Table 2.3: Intercorrelations (Pearson's R) of subscales of GMSI scores.

Results on Objective Music Perception Skills Mini-PROMS categorizes perception skills as "Basic" if the total obtained score is lower than 18, "Good" if between 18 and 22.5, "Excellent" for values between 23 and 27.5, and "Outstanding" for values over 28 [111]. The original Mini-PROMS study covered a total $n = 150$ sample of participants, all recruited from the university of Innsbruck, via email. Most of the participants were students with

	Range	Median	Mean	Standard Deviation(1σ)
Melody	1.5-9	5	4.98	1.59
Tuning	1-7.5	4	4.22	1.62
Accent	0-9.5	5	5.19	1.84
Tempo	1-8	5	5.14	1.59
Mini-PROMS Total	6-30	19.5	19.53	4.98

Table 2.4: Mini-PROMS Range, Median, Mean and Standard Deviation

at least one degree ($n = 134$), aged 27 on average.

We observed (see Table 2.4) an average of "Good" music perception skills for our workers ($\mu = 19.53$, avg. accuracy 54.25%). 48 out of 87 (55.17%) produced reasonably high accuracy in music skill tests (belonging to "Good" and better categories according to Mini-PROMS results). These figures are lower compared to the results of the original study [111] ($\mu = 24.56$, 68.2% avg. accuracy), a fact that we account to the greater representation of *non-musician* in our workers pool (67.82%), compared to the participants of the original Mini-PROMS study (where only 38.67% identified as non-musicians). However, considering the low formal training amongst the surveyed workers, we consider this result an indication of the existence of useful and somewhat abundant auditory music perception skills among untrained workers. Especially, in the top 10% of workers, ranked according to their total Mini-PROMS values, several achieved quite high accuracy, between 73.6% and 83.3%, which would indicate perception skills between "Excellent" and "Outstanding" in Mini-PROMS's scale. In the following section we will analyse in greater detail the relationship between the measured music sophistication and the perception skills.

A similar trend towards lower performance compared to the original Mini-PROMS study can be observed across the other musical aspects: workers correctly identified melody differences with 49.77% avg. accuracy (original study: 64.3%), tuning differences with 52.73% avg. accuracy (original: 68%), accent difference with 51.95% avg. accuracy (original study: 61.5%), and tempo differences with 64.3% avg. accuracy (original study: 81.25%).

The result of the music skill tests is in-line with the result of self-reported music sophistication from GMSI, suggesting that when compared to the populations covered by previous studies, crowd workers generally possess less music perception skills. To deepen the analysis, we calculated the intercorrelation of Mini-PROMS subscales, and made comparison with the original study [111]. Since the Mini-PROMS scores across all the subscales follow normal distributions (Shapiro-Wilk tests[35]), we carried out Pearson's R tests to get the correlation coefficients and corresponding p-values. We find statistical significance on all the intercorrelations. Especially, we find that workers' music skills related to melody are positively correlated with their accent- and tempo-related skills ($R = 0.551$ and $R = 0.514$ respectively), while accent and tempo also shows a moderate correlation ($R = 0.468$). In comparison with the original study, we do not observe large differences in the R values, while we did with the GMSI results. The results of the intercorrelation analysis suggests that worker melody, accent, and tempo skills are related with each other in our population too. This is a positive result, that suggests 1) the applicability of this testing tool also on this population, and 2) the possibility of developing more compact tests for music perception skills, for workers' screening or task assignment purposes.

When focusing on the top 10% of workers, we observed an accuracy on “Melody” between 75% and 90% , while the top 5% scored higher than 85%. A person with “Absolute Pitch” would be expected to achieve high accuracy on this test. Only one person in the top 10% had indicated “Absolute Pitch”, but their accuracy was one of the lowest in the group (75%). This could indicate that the person is more likely to not possess such a characteristic. For the subcategory of “Tuning”, the top 10% achieved accuracy between 81.25% and 93.75%, while the top 5% scored higher than 87.5%. On “Accent”, the top 10% reached accuracy between 80% and 95%. Finally, on the subcategory of “Tempo” we measured accuracy of 87.5% and 100% in the top 10%, while the top 5% achieved perfect score of 100%.

These results suggest the presence of a substantial fraction of workers possessing higher music perception skills than expected from their training, although differently distributed. For example, workers who perceived well changes in “Melody”, didn't perform equally well on the other categories. This could indicate that music perception skills do not necessarily "carry over" from one music feature to the other; other workers will be good in perceiving changes in tempo, while others on tuning. This encourages the use of the appropriate set of tests, to identify potentially high performing annotators. Thus, if we take as example beat tracking annotation tasks, it would be more beneficial to focus on testing the rhythm-related perception skills, as the other categories have lower chance to capture the appropriate workers for the task.

Post-task Survey: Equipment and Cognitive Workload The majority of the workers reported that, during the test, they used headphones (52.87%) (which is very good for musical tasks), earphones (29.54%) and laptop speakers (16.09%) (which are not optimal). All workers reported the quality of their equipment as “Fair” or better quality (55.17% selected “Excellent” and 34.48% “Good”). 96.55% argued that their equipment either does not have any impairment (72.41%) or that the impairment is not annoying (24.13%). Finally, the majority of workers (58.62%) reported near silence conditions, while 31.03% of them reported normal, non-distracting levels of noise. While these conditions are not comparable to lab setups, we consider them to be sufficiently good to accommodate the requirements of our study.

In the NASA-TLX questionnaire, 34.48% of crowd workers reported low “Mental Demand” and 79.31% low “Physical Demand”. “Temporal Demand” was also reported low for the 72.41% of the participants. This low self-reported demand, is reflected also to the majority (55.17%), who reported higher than average “Performance”. Nevertheless, the majority of crowd workers (70.11%) reported average to very high amounts of “Effort” while completing the study, which is not reflected on the perceived mental, physical and temporal demand they experienced. It is also not evident on their “Frustration” levels, since the majority (54.02%) reported low levels.

Using Pearson's R, we found the inter-correlations between the different categories of NASA-TLX. We found high correlation between “Physical Demand” and “Mental Demand”, but also between “Physical Demand” and “Temporal Demand”. Finally, “Frustration” and

"Performance" show high correlation between them, which is a reasonable effect.

Identifying factors influencing performance in Mini-PROMS

To better understand factors affecting a participant's performance in Mini-PROMS and therefore their perceptual capabilities on Melody, Tempo, Tuning and Accent, we applied a Multiple Linear Regression, using Ordinary Least Square (OLS) method. We split our analysis based on total score on Mini-PROMS and the individual categories of the test, to study how they are influenced by the rest of the study's categories.

To minimize multi-colinearity between the Independent Variables, we dropped those that showed high correlation between them in our inter-correlation analysis. Analysing the inter-correlations between all categories, we found similar results to those per part of the study (as analysed in previous sections). Therefore, NASA-TLX was the only part of the study on Prolific, where high inter-correlation was exhibited between the categories of "Physical Demand" and "Mental Demand", "Physical Demand" and "Temporal Demand", "Frustration" and "Performance". We proceeded to apply OLS, by dropping "Physical Demand" and "Frustration" from the NASA-TLX factors, to decrease colinearity. Correspondingly, for the categorical variables "Occupation" and "Equipment Type", we only used the "Part Time", "Voluntary Work", "Unemployed" and "Headphones", "Laptop Speakers" for each respective variable.

For the total Mini-PROMS score, we found a significant equation ($F(19,67) = 2.948, p < .000$, with $R^2 = 0.455$), that shows "Perceptual Abilities" and "Musical Training" from GMSI, affect significantly the dependent variable ($p < 0.05$). For each unit increase reported under the "Perceptual Abilities", a worker showed an increase of 0.2207 point in the total score, while in "Musical Training", it resulted to a 0.2417 increase.

Running the regression for the "Melody" of Mini-PROMS ($F(19,67) = 1.898, p = 0.0289$, with $R^2 = 0.350$), we found that their "Occupation" status affected the dependent variable significantly ($p < 0.05$). Their "Part Time" employment seemed to negatively influence their performance in "Melody" test, by -1.2234 points. On the other hand, "Perceptual Abilities" and "Musical Training" from GMSI also affected significantly their performance ($p < 0.05$), increasing it by 0.0827 and 0.0570 points respectively. Their "Singing Abilities" though, seemed to significantly influence their performance but negatively, where every reported increase on those abilities, resulted to a decrease of -0.0672 point.

The significant regression equation that was found for the "Tuning" category ($F(19,67) = 2.301, p = 0.006$, with $R^2 = 0.395$) showed that their "Occupation" status was yet again affecting their performance significantly ($p < 0.05$). Those who reported "Unemployed" showed an increase in their performance by 0.9205. Finally, "Musical Training" appears to be another significant factor to their performance in this particular audio test. Each unit increase in the category, resulted in a 0.0709 increase in their performance.

For "Accent", the regression ($F(19,67) = 2.580, p = 0.002$, with $R^2 = 0.422$), showed that the "Temporal Demand" the participants experienced, alongside their "Occupation" and "Musical Training", influenced significantly their performance in this test. An increase in "Temporal Demand" resulted in decrease by -0.0851 point and in "Musical Training", an increase by a 0.0701 point. "Part Time" occupation is negatively associated with their performance here, leading to a decrease of -1.2801 points.

Finally, for the "Tempo" category of Mini-PROMS, we couldn't find a significant model by applying OLS.

Amazon Mechanical Turk

Of the 100 workers recruited from Amazon Mechanical Turk (MTurk), 9 of them failed at least one attention check question(s); 7 of them provided invalid/none inputs. After excluding these 16 invalid submissions, we have 84 valid submissions from 84 unique workers.

Worker Demographics We also conducted the same study on Amazon MTurk, in order to see if we can observe similar trends as shown in the last section also on a platform different than Prolific. We gathered 84 crowd workers who provided valid submissions. As seen in Table 2.5, the age range was between 18 and above 65, while the majority was between 26-35 (52.38%), a relatively older pool compared to the Prolific's one. The majority of them were employed (95.23%), with the 88.75% of them full-time. Most of the participants hold a degree (86.90%), with Bachelor's being the most common (60.27%). Finally, the vast majority of the participants, report the United States of America (89.28%) as their residence, with the rest being spread between Brazil (3), India (3), United Kingdom (1), Netherlands (1) and Italy (1).

Apart from education, we see a clear difference between the participants from the two platforms on the age, occupation and country of residence categories. In this study, most of the crowd workers from MTurk are older than those on Prolific, employed and residing in USA.

	Variables	**Statistics**
Age (years)	Range	18-65+
	Majority	26-35 (52.38%)
Occupation	Full-time	71
	Part-time	9
	Unemployed	3
	Retired	1
Education	Associate degree	6
	Bachelor's degree	44
	Doctorate degree	2
	High school/HED	11
	Master's degree	10
	Professional degree	0
	Some college, no diploma	8
	Some high school, no diploma	1
	Technical/trade/vocational training	2

Table 2.5: MTurk participant demographics

Results on Worker Music Sophistication In Table 2.7, we summarise the results of the GMSI questionnaire, regarding the workers on MTurk. As described in 2.2.3, we compare the results on this platform, with the results of the original GMSI study [65].

Comparing our collected data to the original GMSI study, we find that the crowd workers of MTurk exhibit a strongly lower overall music sophistication, at the bottom 32% of the original study. They also score low in all sub-categories, with Musical Training being the only category comparing higher to the 37% of the original study's population.

	Range	Median	Mean	Standard Deviation(1σ)
Active Engagement	12-46	32	30.57	7.92
Perceptual Abilities	18-47	32.5	32.82	5.92
Musical Training	7-43	23	21.80	9.40
Singing Abilities	9-45	32.5	28.29	8.12
Emotions	7-41	30.5	30.34	5.35
General Music Sophistication	29-113	75	72.19	18.15

Table 2.6: GMSI Range, Median, Mean and Standard Deviation

An extremely high number of participants (40.47%), reported having "Absolute Pitch", which is a highly unlikely portion of the sample, as discussed before. Only 9 of them reported adequate formal musical training, which can indicate a general misconception on the entailing traits of such a phenomenon. The reports are much higher than those on Prolific.

With a quick glance at the values on Table 2.6, we see that they indicate skewness on the distributions of each category. When running the Shapiro-Wilk normality test [35], we found that all distributions, except that of "Perceptual Abilities", are non-normal. For that reason, we used Spearman's ranked test to calculate the correlation coefficients between the GMSI sub-categories.

	Active Engagement	Perceptual Abilities	Musical Training	Emotions	Singing Abilities
Active Engagement	1.000				
Perceptual Abilities	0.232*	1.000			
Musical Training	**0.595***	0.263*	1.000		
Emotions	0.213	0.471*	-0.052	1.000	
Singing Abilities	**0.637***	0.340*	**0.552***	0.223*	1.000

Statistical significance of Spearman's rank coefficients, are marked as $^*p < 0.05$.

Table 2.7: Intercorrelations (Spearman's Rank) of subscales of GMSI scores.

We find that "Active Engagement", "Musical Training" and "Singing Abilities" are highly correlated with each other. The positive high correlation between these categories, indicates that crowd workers on MTurk report similarly their aptitude on those GMSI categories. Notably, although not particularly high, there is certainly a positive correlation between self-reported "Perceptual Abilities" and the extent of "Emotions" these crowd workers experience when listening to music ($R = 0.471$).

Results on Objective Music Perception Skills

Table 2.8 shows the results of MTurk's crowd workers on Mini-PROMS test. The mean overall score shows that the average participant in our sample pool, had lower than "Basic" music perception skills overall ($\mu = 15.2$ 42.2% avg. accuracy). This performance is much lower than both the original Mini-PROMS work [111] and the results we retrieved from Prolific. In the Top 10% of the highest performant crowd workers, we see that they score from 61.1% up to 81.94%, scoring from "Good" to "Outstanding", based on the Mini-PROMS scale.

	Range	Median	Mean	Standard Deviation(1σ)
Melody	1-8	4.25	4.22	1.56
Tuning	1-7.5	3	3.2	1.33
Accent	0-7.5	4	4.04	1.42
Tempo	1-8	3.5	3.75	1.64
Mini-PROMS	6.5-29.5	14.5	15.2	4.77

Table 2.8: Mini-PROMS Range, Median, Mean and Standard Deviation

Per individual categories, we see that the highest performance that the crowd workers achieved, didn't reach the max of the Mini-PROMS scale of every category except "Tempo". The avg. accuracy on the "Melody" category, reached 42.2% (original study: 64.3%, Prolific: 49.77%), while on the "Tuning" category, the avg. accuracy was 40% (original study: 68%, Prolific: 52.73%). The participants from MTurk, were able to detect changes on "Accent" features with avg. accuracy of 40.4% (original study: 61.5%, Prolific: 51.95%), while they scored avg. accuracy 46.87% on "Tempo" (original study: 81.25%, Prolific: 64.3%).

Running the Shapiro-Wilk normality test on each Mini-PROMS' category, we find that only the "Melody" one is Gaussian. We used once again Spearman's rank method to calculate the correlation coefficients per category. We found that "Tempo" is highly correlated with "Melody", while "Tuning" is with "Accent". These results are not in line with the original PROMS study [48], but we observe relatively strong correlation between "Tuning"-"Melody" and "Tuning"-"Tempo", which fall into the PROMS categories of "Sound perception" and "Sensory" skills respectively.

Post-task Survey: Equipment and Cognitive Workload The majority of the participants from MTurk used headphones to perform Mini-PROMS (64.28%), while 20.23% used earphones and 15.46% used the speakers of their laptops. Most participants described the condition of their equipment as "Excellent" or "Good", while one reported it as "Fair". The majority (66.66%) of the crowd workers, reported any impairment of their equipment as "Impairceptible", with 15.47% of them describing it as "Perceptible but not annoying". The rest of the workers reported various degrees of annoying impairments. Finally, 65.47% of the crowd workers performed Mini-PROMS with near silence environmental conditions, while 19.04% reported extreme levels of noise around them. None of the distributions passed the Shapiro-Wilk normality test for each equipment-related category. Running Spearman's rank method, we found no notable correlation between the categories.

In the NASA-TLX questionnaire, 29.76% of crowd workers reported average "Mental Demand", with 10.71% and 15.67% reporting low or very high mental strain respectively. 46.43% reported low "Physical Demand" with 36.9% of the total not feeling rushed while performing the study. 27.38% reported average "Performance", with 22.61% describing their performance as successful. The majority of participants were divided between reporting high effort (27.38%) or moderate difficulty (27.38%). Finally, 34.52% of the crowd workers felt little to no frustration with 20.24% reporting moderate levels.

Using Spearman's rank, we found that "Frustration" is highly correlated with "Physical Demand", "Temporal Demand" and "Performance". This shows that the more physical strain and hurried they felt, combined with feelings of failing the task at hand, increased their frustration with the study.

Identifying factors influencing performance in Mini-PROMS

Following the analysis on the results from Prolific, we applied Multiple Linear Regression on the Mini-PROMS categories, using the Ordinary Least Square (OLS) method. For the total Mini-PROMS score as the dependent variable ($F(18,65) = 4.742, p < .000$, with $R^2 = 0.567$), we found that only the "Perceptual Skills" from GMSI and the "Physical Demand" category from NASA-TLX, affect significantly the dependent variable ($p < 0.05$). For each extra point reported under the "Perceptual Skills", a worker showed an increase of 0.2 points in the total score. On the other hand, a single extra point towards "Very demanding" on the "Physical Demand" category, resulted on a -0.3 decrease of total performance by the worker.

Running the regression for the "Melody" of Mini-PROMS ($F(18,65) = 3.443, p < .000$, with $R^2 = 0.488$), we found that "Physical Demand" category from NASA-TLX affected the dependent variable the most ($p < 0.05$). The effect is negative towards the performance on "Melody", where each point increase on "Physical Demand" translated to -0.12 point decrease of performance.

The significant regression equation that was found for the "Tuning" category ($F(18,65) = 1.849, p = 0.0376$, with $R^2 = 0.339$) showed that the most significant factor was yet again the "Physical Demand". The more physically demanding the study was perceived, it influenced the final score on "Tuning" by -0.09.

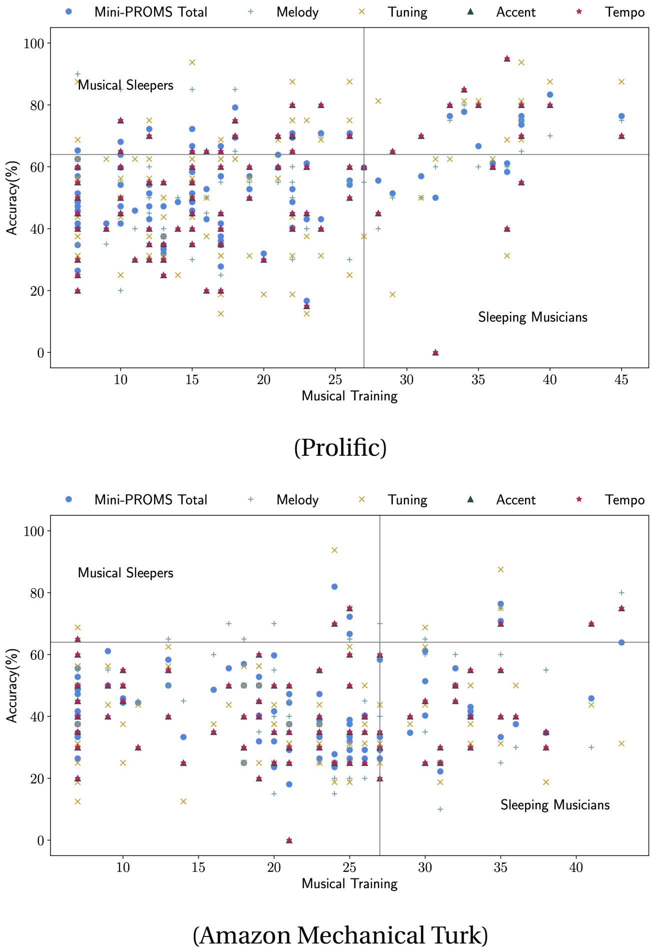
For "Accent", the regression ($F(18,65) = 2.130, p = 0.0141$, with $R^2 = 0.371$), showed that the "Type" of audio equipment and its "Impairment" affected the workers' performance the most. The "Laptop Speakers" seemed to influenced positively their performance by 1.2193 points, while the less perceptible an "Impairment" was, it was increasing their performance by 0.448 point.

Finally, for the "Tempo" category of Mini-PROMS, we found a significant regression equation ($F(18,65) = 4.502, p < .000$, with $R^2 = 0.555$) that shows that "Physical Demand" and "Perceptual Abilities" influenced the performance on the category most significantly. While an increase in "Physical Demand" decreased the performance by -0.10 point, an increase in the self-reported "Perceptual Abilities" showed an increase of performance on "Tempo" by 0.08.

In search of Musical Sleepers

Having analysed each component of our study and using OLS to understand how individual factors could have influenced the workers' performance on the perception skills of Mini-PROMS, we were still interested to investigate how the highly perceptive workers are distributed based on quantifiable expertise. Musical training is an element that can be quantified by questions on credentials, years of education etc, all components that can be retrieved by the respective category in GMSI. It is an attribute that we experts show high proficiency and that a platform could potentially easily store and iterate per worker's profile.

In this study, following the original studies of PROMS [48] and Mini-PROMS [111], we make the comparisons of levels of Musical Training, against the performance on the categories of Mini-PROMS. Taking a step further, we used as baselines the amount of “Musical Training” that 50% of the original GMSI’s population exhibited (27) and the lowest bound of “Excellent” performance (63.98%) on perception skills, as established for Mini-PROMS. We make use of the terms “Musical Sleepers”, to label those who exhibit high performance but reported low training and “Sleeping Musicians”, those who reported extensive training but performed poorly, both terms from [48] and [111].



(Prolific)

(Amazon Mechanical Turk)

Figure 2.4: Musical Training (GMSI) and Performance on Mini-PROMS (acc%).

Figure 2.4 shows a scatter plot per platform, that shows how participants are distributed based on their performance and “Musical Training”. We witness that on both platforms, there is a high number of crowd workers who reported low “Musical Training” (below 50% of original GMSI study [65]) and had relatively low performance in the Mini-PROMS tests. This is to be expected, due to the nature of the domain and the niche skills that are required.

The attention is naturally drawn to the “Musical Sleepers”; a portion of the population that can exhibit relatively high music perception skills, but did not have adequate education. Few people would follow any form of dedicated music studies, making it even more difficult to find them on a crowdsourcing platform. With low expertise being the norm, finding crowd workers with high, untrained, auditory skills among them, is a rare phenomenon that could greatly benefit systems who would make use of such skills. In the case of Prolific, we witness “Musical Sleepers” in a higher number compared to Amazon Mechanical Turk. We cannot draw platform-based conclusions though, since our participant pool was quite small relatively to the actual population of each platform. The presence of these workers is very encouraging, as it shows that it is possible to deploy advanced music analysis tasks on microtask platforms and and finding high-value contributors.

In our study, participants from Amazon Mechanical Turk, generally reached lower performance compared to the ones from Prolific. This is an outcome also evident on the high number of “Sleeping Musicians” on MTurk, compared to the smaller portion of the total Prolific participants. These workers reported relatively high musical training, but performed lower than expected from a person of their expertise.

2.2.4. Discussion

In this study, we extensively measure the musical sophistication and music perception skills of crowd workers on Prolific and Amazon Mechanical Turk. We show that on both platforms, the self-reported music sophistication of crowd workers is below that of the general population and that formally-trained workers are rare. Nevertheless, we found surprisingly refined and diverse music perception skills amongst the top performers per platform. These skills though cannot accurately and easily be predicted by questions.

On music perceptual skills and predictors

Workers on both platforms exhibited quite diverse set of music perception skills. Among the high performant ones, we found evidence that supports the existence of workers with high accuracy and little to no formal training, namely “Musical Sleepers”, indicating the prospect of high-quality annotations by non-experts on these platforms. Predicting these skills though, can prove far from trivial. To promote reproducibility of our results, we made use of established tools to retrieve domain sophistication [65], perceptual skills [111], perceived workload [36], equipment condition [77] and ambient noise levels [9].

In an analysis of workers’ reports on other parts of the study, we found per platform, different factors that significantly correlated to their performance. “Musical Training”, a type of expertise that could be thought as a strong indicator of a worker’s perceptual skills, showed low significance on the performance of Amazon Mechanical Turk workers. These findings, alongside the high number of “Sleeping Musicians” among the participants from Amazon Mechanical Turk, indicate a notable difference between their reported knowledge and their quantified perceptual skills. On the other hand though, the self-reported “Perceptual Abilities” proved a reliable factor of MTurk workers, as they were significantly related to their performance on Mini-PROMS. This is in contrast to the reported “Perceptual Abilities” of Prolific’s workers, which did not significantly correlate to their performance. Aspects of perceived task workload though, as retrieved from NASA-TLX, seemed to significantly correlate on categories of the Mini-PROMS test, on both platforms. Finally, while demographic data appear relevant to aspects of the performance of workers on Prolific, on MTurk equipment showed to play a more important role on the “Accent” test of Mini-PROMS.

The “Active Engagement” category of GMSI, which indicates to what extent a person engages with music as a hobby (frequenting online forums, buying music albums etc), did not show any significant correlation to the measured music perception skills of the participants on both platforms. That shows that we cannot reliably use such questions, to infer the skills of the worker; the time/effort spent listening to or discussing about music, can be indifferent of the range of their skills. The same applies to the “Emotions” category, where participants report their emotional response to music. This indicates that music could still evoke emotions to people, even without them perceiving its structural elements.

Implications for Design

Self-reported Musical Sophistication. The musical sophistication assessments (GMSI) is a useful tool to evaluate workers’ capability in completing music-related tasks. It is however a lengthy questionnaire, which could result in extra cost and worse worker engagement. Reducing the number of question is possible, but with implication in terms of test reliability. For instance, the subscale of Musical Training is positively correlated to their actual music perception skills (and the correlation coefficient is higher than the general GMSI). As music perception skills are of primary relevance when executing music-related tasks, we suggest that in future task design, requesters could consider using the subscale of musical training which only contains 7 items. This could be complemented with novel methods to effectively and precisely predict worker performance to further facilitate task scheduling and assignment.

Music Perception Skill Assessment. The Mini-PROMS tool appears to be an effective mean to evaluate worker quality in terms of music skills. Yet, it suffers from the same overhead issues of GMSI. In this case, we suggest to use PROMS or Mini-PROMS as a qualification test, possibly featured by crowdsourcing platforms. Workers could use this test to get the corresponding qualification, to obtain the opportunities to access more tasks, and earn more rewards.

Music Annotation and Analysis Tasks. The results of this study indicate that knowledge- and skill-intensive musical tasks could be deployed on microtasks crowdsourcing platforms, with good expectations in terms of availability of skilled workers. However, performance on different skills (Melody, Tuning, Accent, and Tempo) appears to be unevenly distributed. We therefore recommend to analyse the capabilities of the selected crowd and tailor the design of advanced music annotation and analysis tasks to precise music perception skills.

Limitations and Future Work

A main limitation of our study is concerned with the size of the tested population. While we employed workers from two different platforms, our results cannot be generalised per platform. A larger participation pool could potentially aid the generalisability of our findings and lead to more fine-grained insights. Even though our results are based on a population of crowd workers that have received less formal musical training than the average population used in similar studies, [65], the use of standardised and validated tests, lend confidence to the reliability of our findings.

Another potential confounding factor in our study is the motivation for participation. We attracted crowd workers using monetary rewards, while in other studies people voluntarily performed their test (e.g. BBC's main Science webpage) [65]. Such a difference could also explain the differences in observed distributions (musical training and perception skills). However, monetary incentives are a feature of crowdsourcing markets, which makes them appealing in terms of work capacity and likelihood of speedy completion. In that respect, our findings are very encouraging, as they show the availability of both musically educated and/or naturally skilled workers that could take on musically complex tasks.

As demonstrated in our results, workers who perform well in a certain perception category (e.g. “Melody”) do not perform equally well in another (e.g. “Tempo”). In future studies, we encourage the use of perception tests, adjusted and adapted for the specific music task at hand by using the appropriate categories, to accurately select potentially highly performing workers.

In our analysis, we currently made use of Ordinary Least Square Regression to identify factors are associated with the workers' performance on Mini-PROMS. Although this method gave us some first insights, further studies are needed to expand our pool of crowd workers and use other models that can help us find predictors of perceptual skills of workers accurately. This could assist in designing appropriate task assignment methods, to increase the efficiency and effectiveness of crowdsourcing systems that make use of such skills.

In this study, we utilized standardized tools to capture domain-specific characteristics of the workers of a specific platform. Comparing results from their self-reported "connection" to the domain, with those from actively testing their skills, can paint a clear picture of the workers' demographics on a specific domain. While this work is specific to the music domain, we believe that similar workflows can be utilized to study the characteristics of workers on other domains. This holds especially true, as crowdsourcing platforms have diverse user-bases and direct comparisons cannot safely be drawn to studies with highly controlled population samples.

2.2.5. Conclusion

To summarize, in this section we presented a study exploring the prevalence and distribution of music perception skills of the general crowd in the open crowdsourcing marketplace of Prolific and Amazon Mechanical Turk. We measured and compared self-reported musical sophistication and active music perception skills of crowd workers by leveraging the established GMSI questionnaire and Mini-PROMS audio-based test respectively. Our analysis shows that self-reported musical sophistication of crowd workers is generally below that of the general population and the majority of them have not received any form of formal training. We observed differences in the two participant pools, on both their performance and factors which are significantly correlated to it. Nevertheless, we identified the presence of *musical sleepers* on both platforms. Moreover, our analysis shows worker accessibility to adequate equipment. Together, these findings indicate the possibility of further increasing the adoption of crowdsourcing as a viable means to perform complex music-related tasks.

2.3. User Engagement with Classical Music on Online Platforms

In the general public, the popularity of classical music composers is often approximated through commercial figures like album releases, record sales, or live performances. However, commercial factors only provide one piece of the overall picture. The success of community-driven platforms has profoundly changed how people consume and interact with music, and, consequently, our understanding of what popularity is. People discuss their favourite artists, archive knowledge regarding them and share their work through multimedia platforms. This leads to knowledge creation and propagation that is often overlooked when preserving classical music. As a first step, in this section, we investigate how data from these platforms can provide a more comprehensive view on popularity and engagement regarding the long-tail of classical music composers.

Popularity is a desired achievement for artists, as it promotes their work, facilitates the interaction with the broader audience, and ultimately affects how people will remember them in years to come. This holds especially true in classical music, which includes centuries-long catalog of works created and reinterpreted by a long list of artists. Access to those works and their artists is essential in preserving parts of historic and cultural heritage, as classical music has fundamentally influenced western music throughout history.

Recording labels have pioneered the methods to capture audio performances, preserving a plethora of classical music pieces of the old or contemporary composers. Inevitably though, these recordings are skewed towards already established and well studied composers and their works can be found performed by various artists. However, popularity is not only related to talent, but also former success [57] and "*the need of consumers to consume the same art as others*" [1]. When people are exposed to art, share it with others, or consume media about it, they create what is called "consumption capita" [99]. These activities have become much easier to perform with the widespread adoption of Web technologies and the availability of community-driven platforms. People from around the world with different cultural backgrounds can openly share their knowledge, experiences, and recordings with others in the ever expanding publicly-accessible knowledge bases and social media communities.

For this reason we believe that by tapping into user-generated content, we can find interesting insights on user-engagement and popularity of classical composers online. To that end, we analyze user-generated data on Wikipedia and YouTube compared to album releases retrieved from MusicBrainz[9], to investigate our research question:

[9] `https://musicbrainz.org/`

- **[RQ4]** How does the popularity of classical music composers on community-driven platforms relate to their album release trends?

Our findings can be used to indicate to what extent those platforms potentially hold content and information about composers who have zero to low number of official album releases. Results also show that enthusiasts exist in such online communities, who could be potentially be recruited in music annotation projects. These enthusiasts, through their online engagement, exhibit signs of musical affinity and self-motivation to preserve and share classical music information; valuable characteristics for music annotators.

2.3.1. Related Work

Research on the multilingual online encyclopedia Wikipedia[10], has shown that it is possible to predict real-world opinions and popularity through analyzing user interaction on the platform. For example, Mestyán et al [64] presented a model using Wikipedia's user-generated content that accurately predicts movie box office success, while another study by Wei et al [107] uses user interaction on Wikipedia to predict stock market values.

YouTube[11] is a video-sharing, social media platform, which encourages a user-content-user interaction [106] and hosts a staggering amount of videos across a wide variety of categories. While research work has mostly focused on network dynamics [96, 73, 95] and opinion mining [71, 94, 93], there have been studies on popularity on the platform. In Chatzopoulou et al. is shown that views, number of comments and ratings are highly correlated with each other, while in [13] it is found that popularity on YouTube follows the results of studies on how "consumption capital" is increased, as seen in [2] (i.e. "consuming" more content of creators, increases our enjoyment towards them and their art [99]). Related to our study, the work of Cayari [17] studies how YouTube has fundamentally affected musical art forms and has essentially changed the way people listen, share and consume music. This holds especially true for classical music, where with a quick search on the platform, it is possible to find several results with multiple compilations of classical music works, live footage of concerts, interpretations of compositions, and educational material.

Our study is inspired by the work of Bellogin et al [10], the first comparing music artist popularity from different Web and music services. The sources used in that study were domain specific services (namely EchoNest[12], Last.fm[13] and Spotify[14]), looking into music consumption trends. The main difference with our work is that we look into user-engagement regarding music on generic platforms that host music information among other categories. In our study, we also focus on genre-specific content on those generic platforms, rather than overall music consumption trends. Related to classical music, Schedl et al. [91] studied the online engagement of fans of classical music on Twitter[15] and Last.fm[16] finding that classical music is under-represented on social media, as classical

[10] https://en.wikipedia.org/wiki/Wikipedia

[11] https://www.youtube.com/

[12] the.echonest.com

[13] https://www.last.fm/

[14] https://www.spotify.com/

[15] https://twitter.com

[16] https://www.last.fm

music enthusiasts seem averse sharing content on both of the studied platforms. In this study we argue that we need further analysis on community-driven platforms to be able to assess the extent those platforms contain domain specific information.

2.3.2. Data Collection

We started with quantifying user engagement on Wikipedia and YouTube. We defined a list of classical music composers which we are considering for analysis in different platforms and selected a reliable source of knowledge about album releases. In this section, we outline our data collection process. The data retrieval and scripts, as well as the resulting dataset are available here[17].

DBpedia: to select the composers, we retrieved a list of their names from various classical music periods from DBpedia[18] using the Virtuoso SPARQL Query Interface[19]. We used the name contained in the URI as an unambiguous representation of a composers' name, as there could be multiple entries with the same name, but different people. We queried the composer names, using the following Yago entities available for classical music periods in DBpedia:

- WikicatBaroqueComposers
- WikicatClassicalComposers
- WikicatClassical-periodComposers
- WikicatRomanticComposers
- Wikicat18th-centuryClassicalComposer
- Wikicat19th-centuryClassicalComposers
- Wikicat20th-centuryClassicalComposers
- Wikicat21st-centuryClassicalComposers

We finally collected 5928 different composers, distributed as follows: 1126 baroque, 1025 romantic, 725 classical, 96 from eighteenth century, 345 from nineteenth century, 3155 from twentieth century and 1383 from twenty first century. Obviously, a composer can belong to multiple periods.

Wikipedia: we used Wikipedia APIs[20] to retrieve each composer's page and related information. To ensure we retrieve the page related to the composer, we use as query the name as it appears in the "About" on DBPedia. For instance, to retrieve the page of *Alexander Müller* we use *Alexander_Müller_(composer)*. This also holds true for composers having the same name, like *Johann Strauss I* and his son *Johann Strauss II*. We follow the same procedure for a composer's full name on MusicBrainz and YouTube. More specifically, we retrieved: 1) the number of edits a page has received, 2) the number of users who edited the page, 3) the number of languages the page is translated to, 4) the size of the page (in KB), and 5) how many sections it contains. We use these data as a proxy to measure how many people engage on Wikipedia with a composer's entry. They show us the amount of users maintaining a page, how much are they committed in keeping the information correct and updated (number of edits), accessible to many people (languages) and complete (page length and number of sections).

[17] `https://github.com/ipsamiotis/classical_popularity`

[18] `https://wiki.dbpedia.org/`

[19] `https://dbpedia.org/sparql`

[20] `https://www.mediawiki.org/wiki/API:Main_page`

MusicBrainz: Musicbrainz is a community-maintained music encyclopedia that collects metadata about music artists and their released works (albums, recordings and more). We used their up-to-date and proven reliable information on *Album Releases* per classical music composer. We retrieved this using a Python wrapper to its API[21] and in our analysis we refer to them as *Album Releases*.

YouTube: we gathered videos from YouTube using youtube-dl[22] based on the queries: `"composer name" + "music"` and `"composer name" + "live music"`, resulting in a total of 184,019 video entries. The retrieved videos contained a high amount of non-relevant entries. We cleaned the dataset by discarding all videos where the composer name was not present in either the title, in one of the tags or in the description of the video. This way, we decreased the number of videos relevant to our study to 69,261. Since users also view videos and react to them by leaving a like, dislike or a comment, we also retrieved the number of: 1) likes, 2) dislikes, 3) views, 4) comments, 5) unique uploaders and 6) duration of videos. From these engagement data, we further calculated the average number of likes, dislikes, views, video duration and comments per composer.

In the final dataset we collected, we find 97% of the composers have a Wikipedia page, 87% of the composers have at least one video on Youtube, while only 70% of them have albums on MusicBrainz. Also, 63% of composers are in common to all the platforms, with Wikipedia and YouTube having the biggest overlap (99%), while MuiscBrainz lacks 31% of composers present in Wikipedia.

2.3.3. RESULTS

We first analyse if the data we use as a proxy for user engagement on Wikipedia and YouTube is correlated. To that end, we calculated the Spearman correlation [66] of the popularity ranking of composers each data property was generating. This will give some insights into how comparable user engagement is between platforms. To investigate if user-engagement-driven popularity exhibit differences compared to the album releases, we calculate the similarity between those trends. Using the Jaccard similarity index [39], we are able to observe differences between the different popularity rankings and album releases, finding interesting results on both top-to-bottom and bottom-to-top rankings. The intuition is that by computing the Jaccard index at different level of cut-off it is possible to see how much each platform agrees on who are the top *n* (or bottom *n*) composers.

CALCULATING POPULARITY PER PLATFORM

We calculate a single popularity ranking per platform based on user activity metrics as described in Section 2.3.2. These metrics are related to: 1) how much the people engage with the platform in the context of classical music and 2) what kind of data they create. Their connection to popularity is based on the fact that people who are interested in a composer and are active users of those platforms, they will either create data about them, interact with data others created and/or discuss about them with other users. All these factors are an “online version” of those described on “consumption capital”, which indicates the reasons why artists become famous among people, apart from their

[21] `https://python-musicbrainzngs.readthedocs.io/en/v0.7.1/`

[22] `https://pypi.org/project/youtube_dl/`

talent [99]. We first calculate the correlation of those metrics, following the methodology of [18], to find if the engagement metrics we chose are correlated with each other, to be later considered part of a platform-wide “popularity ranking”. This popularity ranking, as discussed before, is being used as a proxy of real popularity of composers on those platforms.

Wikipedia: we compute the correlation of the the metrics described in Section 2.3.2 using the Spearman coefficient. As shown in Table 2.9, most of the different metrics exhibit high correlation (greater than 0.5), with the least correlated being the number of languages with the page size and number of sections metadata.

	Wikipedia				
	Sections	Languages	Revisions	Users	
Pages Size	0.82	0.38	0.72	0.69	
Sections	1	0.38	0.63	0.61	
Languages		1	0.49	0.67	
Revisions			1	0.87	
	Youtube				
	Views	Likes	Dislikes	Duration	Comments
Videos	0.86	0.85	0.81	0.89	0.05
Views	1	0.98	0.94	0.81	0.07
Likes		1	0.93	0.81	0.07
Dislikes			1	0.76	0.06
Duration				1	0.06

Table 2.9: Spearman correlation the metrics of Wikipedia and Youtube

Youtube: we performed the same analysis on the data gathered from YouTube, finding that they are all highly correlated - as shown in Table 2.9 - except the number of comments. This could be explained by limitations on the number of comments which could be accessed by the APIs, so we excluded them from the final rankings. The others are consistent with the results reported in [18].

To compare the popularity of composers on Wikipedia, YouTube and album releases, we needed a single ranking per platform. To achieve this, we ranked the composers based on each metric per platform. If a composer is not present on a platform, the values of his or her platform metrics are set to 0 (e.g., for composers not present on MusicBrainz, we set their album releases number to 0). In this way we obtain list of the same length and we can compute correlation coefficients. Then, we aggregated the different rankings into a single one for each platform using the average ranking aggregation method. We then compared them using the Spearman correlation coefficient.

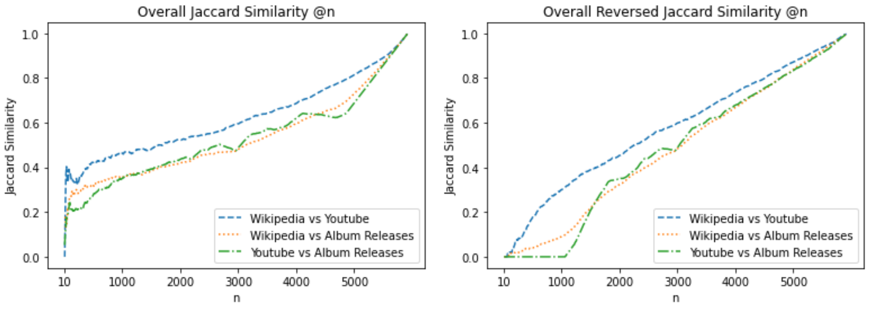
	Wikipedia	MuiscBrainz
YouTube	0.42	0.34
Wikipedia	1	0.36

Table 2.10: Spearman correlation between the platform rankings and album releases.

As Table 2.10 shows, all platforms are positively correlated, with YouTube and Wikpedia exhibiting a stronger correlation to each other's popularity rankings.

Popularity on Platforms and Album Releases

We then calculate the similarity between the popularity rankings derived from the community-driven platforms and the rankings obtained by looking at composers' *Album Releases*. Such comparison can inform us on the degree to which online communities engage in data creation, compared to the official recordings produced by the industry shown in the *Album Releases*. We compute similarity at different ranks, which intuitively can be encoded as "how much different rankings agree" with each other. To that end, we used the Jaccard similarity index. We observe - in Figure 2.5a - that Wikipedia's and YouTube's popularity rankings are more similar to each other, than compared to the *Album Releases*. While this was expected, considering the fact that they are both community-driven platforms (see Table 2.10), this is still notable as both platforms have strong differences in scope and purpose. Wikipedia's ranking is closer to those of *Album Releases*, when compared to YouTube. This results actually is more in line with related work in [64, 107], that shown user-engagement metrics can predict real-world popularity to a certain extent.



(a) (b)

Figure 2.5: Jaccard similarity at different ranking cut-offs with normal (a) and reverse ranking (b)

We compute the Jaccard similarity also by looking at the bottom of the ranking, that is by looking at the long-tail of each platform. Following a reverse ordering on popularity rankings, starting from most "obscure" composers, the uniqueness of YouTube's ranking becomes more evident, as we see in Figure 2.5b. The popularity of the long-tail composers on YouTube is completely dissimilar to the *Album Releases*, as it stays at 0 until we consider 1000 composers. This dissimilarity means that for composers with 0 released albums, the users still engage in data creation on YouTube. This is very encouraging for future works that need to find information regarding classical music composers, as YouTube contains video artifacts and user-engagement metadata for those composers who don't have registries with official album releases. In the same way, Wikipedia, noticeably less severely, has low similarity with *Album Releases* as well for the first 1000 composers. This means that even composers with low-count of official album releases, there are still Wikipedia entries with stored information, generated in its entirety by the community of the platform. These findings strengthen our hypothesis that community-driven platforms behave similarly to each other and differently to *Album Releases*, potentially holding user-generated information for composers with low number of official album releases.

2.3.4. DISCUSSION & CONCLUSION

In this work, we investigated to what extent popularity of classical music composers on generic community-driven online platforms, follows official album releases as registered in MusicBrainz. We found that Wikipedia and YouTube although they share similarities with each other regarding composers' popularity rankings, they differ on the long-tail of popularity, based on album releases. We discovered that Wikipedia's popularity ranking follows more closely the ranking found in album releases, following results from similar studies on the platform. This comes in contrast with the popularity rankings as witnessed on YouTube, which don't follow as closely those of album releases. For example on YouTube, there are many composers who have no entry for an album release on MusicBrainz but they still garner a notable number of followers who engage with their work. This could reflect the more democratized and diverse manner of information creation on YouTube, compared to the more rule-based and expert-driven one on Wikipedia. Due to the similarity of rankings between Wikipedia and YouTube, we find that complementary studies of different platforms could assist to a more holistic overview of published corpora, especially in classical music. Therefore, community-driven online platforms show potential in preserving information regarding composers and their works that are under-represented in the official recorded canon.

Such platforms also show potential as sources of motivated and enthusiastic participants, who exhibit familiarity with (or even knowledge of) concepts of classical music. In the future, we encourage to approach potential annotators through such platforms [43], and understand the relation of their online presence with their perceptive abilities, to complement their music affinity profiles.

3

Designing Crowd-Enhanced Annotation Systems for Music Compositions

Simultaneously to conducting research on the musical characteristics and profiles of people online, we also had to tackle the design and engineering challenges that arise during the development of hybrid transcription workflows. While information on the music affinity of annotators can enable better inclusion of humans in the transcription processes, it does not guarantee their performance on specific types of transcription tasks. In this chapter we explore different possible tasks that can be vital in hybrid transcription workflows (identifying instruments from audio and spotting errors in transcription), while we also present our very own design and implementation of a semi-automatic, collaborative transcription system. The outcomes of these designs and implementations show once again the versatility of non-experts in music transcription tasks, while also underlining the needs and requirements of semi-experts in developing human-assisted music transcription systems.

This chapter is based on excerpts published in deliverables during the European H2020 project ***T****owards* ***R****icher* ***O****nline* ***M****usic* ***P****ublic-domain* ***A****rchives, and the following publications: "Scriptoria: A Crowd-powered Music Transcription System" [82], "Microtask Crowdsourcing for Music Score Transcriptions: An Experiment with Error Detection" [88] and "Crowd's Performance on Temporal Activity Detection of Musical Instruments in Polyphonic Music" [84]*

3.1. Design, Deployment and Evaluation of Scriptoria

Inspired by recent works in microtask crowdsourcing [88, 12] and *nichesourcing* [69], we conducted requirement analysis with experts and designed Scriptoria, a crowd-powered music transcription system. The system consists of multiple modules which process incoming PDF files of scanned scores, process them and segment them into smaller parts. Each segment is then annotated through a crowdsourcing pipeline of incremental transcription (i.e., step-by-step conversion of those image segments to machine-readable format). The results are then aggregated and published in an online repository.

We evaluated our system with Focus Group Discussions with members of Dutch youth orchestras in two iterations. Between these iterations, we made improvements to our system based on the feedback we received. In this paper, we finally present some of the most valuable insights gathered through our discussions with the participants.

Our work has parallels to studies such as [14], the *Allegro* system, and [20], which focused on user input and task design. However, both studies focus on single-user transcription, while our workflows allow many contributors to participate in parallel and on separate segments, to enable better scalability [83]. It is worth mentioning that contemporary pilot studies on performance annotation (e.g. the CosmoNote system [24]) have shown encouraging results on utilizing expert and semi-expert annotators in web-based music annotation systems. Nevertheless, the modalities of their inclusion in human-in-the-loop workflows and task design approaches for transcribing scanned score sheets into machine-readable formats remains an open question that we tackle in this study.

We present Scriptoria in two parts: (a) the back-end architecture of the transcription pipeline and processing modules; and (b) the task interfaces that users interact with. Both the back-end and front-end of this crowdsourcing system are hosted on the Dutch national e-infrastructure with the support of SURF Cooperative. Their source code is published on GitHub[1][2].

3.1.1. System Architecture

Our back-end accommodates a crowd-assisted OMR pipeline. It processes PDF input data (image processing and segmentation), generates crowdsourcing tasks for the non-automated parts, and finally aggregates results to build an MEI version of the original orchestral music score (see Figure 3.1).

The human computation aspects of Scriptoria's backend follow research and findings presented in previous chapters, breaking down the problem of OMR into steps and tasks that can be performed by people on specific stages of the OMR process. The crowdsourcing tasks of Scriptoria have specific inputs (segments of the given score) and outputs (annotations), designed to be performed easily and efficiently by the users. These crowdsourcing tasks co-exist with automated methods such as measure detection, image segmentation and XML-tree aggregation, creating a hybrid system where human-machine collaboration achieves the shared goal of generating MEI orchestral pieces from PDF input.

[1] https://github.com/cakefm/crowd_task_manager

[2] https://github.com/cakefm/scriptoria

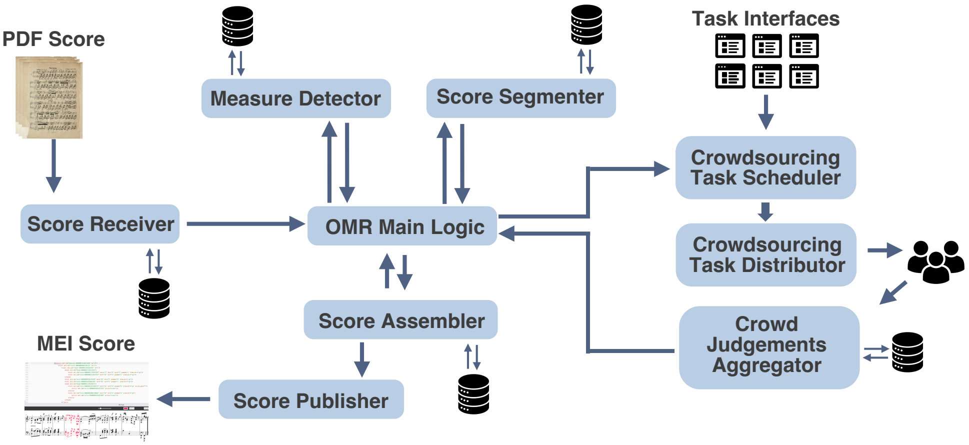


Figure 3.1: Schema of Crowd Task Manager's modules.

Core system requirements for our prototype were to: (1) design the system in a modular and distributed fashion and (2) store in the system all the data resulted from processes throughout each of the steps of our music transcription pipeline, to make them easily accessible by all the system's modules. We set the first requirement to enable scalability and support easier maintainability. Each of the modules in the prototype represents a step on the transcription pipeline and serves a specific functionality. This helps to easily replace parts of the pipeline with more sophisticated ones, without breaking the overall operability of the system. We implemented a central module which holds the logic steps of the transcription pipeline which sends messages that dictate which of the modules should be activated and when. Each module inside the pipeline, imports data from our local database and stores data to it to make them available to the other modules.

When a PDF score is sent to the back-end, first rasterisation is performed followed by some standard image pre-processing. First, the contrast of the page is maximized, after which the page is binarized. Following this, any rotations in the page that might have occurred due to scanning of the original score are rectified. Following these steps, a top-down approach is followed in analyzing the page structure. First, systems will be separated, which are subsequently segmented into vertical blocks, which are then segmented into measures. As the extraction process may not be 100% accurate, this procedure requires a human (expert) post-check before proceeding. Each segment is stored in a MongoDB database, alongside their identifiers and they become available to the front-end through an API.

Consulting an expert from the Royal Concertgebouw Orchestra[3] (RCO) of Amsterdam, we identified the most important elements of an orchestral music score, alongside the minimum-viable-product requirements for a final transcribed score. The music notations we focus in our transcription pipeline where: (a) clefs, (b) time signatures, (c) key signatures, (d) rhythmic information of notes, (e) pitch information of notes. We then broke down the transcription pipeline into consecutive tasks that can be conveyed in a micro-

[3] https://www.concertgebouworkest.nl/en

task crowdsourcing fashion, focusing on those individual important music elements. The tasks we designed were:

- **Clef recognition**: indicate whether one or multiple clefs are visible in a segment, and if so, which;
- **Time signature recognition**: indicate whether a time signature indication is visible in a segment, and if so, which;
- **Key recognition**: indicate whether a key signature indication is visible in a segment, and if so, which;
- **Rhythm transcription**: transcribe the rhythm of the musical content in a segment;
- **Pitch transcription**: transcribe the pitches of the musical content in a segment.

As contributors work on the tasks, in the back-end, a score is built up in the open Music Encoding Initiative (MEI) format, and each contribution is stored as a commit on GitHub. To allow for coherent completion, we implemented a revised scheduling algorithm for Scriptoria's backend, which follows a hierarchy of importance for MEI elements. For each measure, the clef, key and time elements are essential, as they could alter all subsequent music elements (notes/rests) that depend on them. That way, even in the case that a campaign would not fully manage to conclude, Scriptoria's processes will still result to an output that is semantically coherent and comprehensive, rather than having contributions at random parts of a large score.

Scriptoria allows the crowdsourcing tasks to co-exist with automated methods (measure detection, image segmentation, task scheduling and XML-tree aggregation), creating a hybrid system where human-machine collaboration achieves the shared goal of generating MEI orchestral pieces from PDF input.

3.1.2. TASK DESIGN AND INTERFACES

A dedicated front-end server was developed, to allow dynamic rendering of UI elements and dynamic route matching for the different types of tasks. This is based on a NodeJS server which hosts all the necessary components such as interfaces, UI elements and dedicated task type components, while handling communications with the back-end through Axios. The front-end can access each score segment through the back-end's API and renders their images dynamically on the browser. The user input is translated to MEI headers and communicated back to the back-end.

Our target contributors were semi-experts (youth/student members of classical orchestras), so we factored their expertise during the task design process. As described in Section 3.1.1, we designed a separate task type for each of the five music notations. The detection tasks for **clef**, **time signature** and **key signature**, presented the user with the original segment of the score (an image of a given measure) and they had to indicate the existence of the given music notation, while identifying its characteristics (e.g. if a clef exists, select its type). For the **rhythm** and **pitch** detection tasks, the user was presented with the image of the original segment to the left, so they can immediately compare their choices on the rendered MEI snippet to the right (see Figure 3.2).

Due to our contributors' expertise, a certain high level of input was expected. Their expertise, combined with majority voting aggregation and tree aligning algorithms, would ensure high quality of output, therefore rendering possible verification tasks inessential.

3.1.3. Iterative Design and Evaluation

We conducted focus group discussions with semi-experts and young professional members of multiple classical youth orchestras to evaluate our workflow and task designs. During the discussions, we investigated current methods they employ to transcribe orchestral music scores, but also encouraged them to explore the requirements and workflows of a future, more feature-rich version.

Interviews with experts played a crucial role to the design of our transcription system. The initial design of our prototype was based on feedback and requirements received by a professional expert in the RCO and from the youth orchestra Krashna Musika[4]. For the focus group discussions, we reached out to several youth orchestras in Netherlands, who were enthusiastic to participate. Following an iterative design methodology, we split the participants into two large groups; the feedback received from the first group was used to update the designs in our transcription system and the second group was presented with the final version.

All study sessions consider the first pages of Ludwig van Beethoven's Sextet in E-flat major, op. 71, where for the scanned score we used a PDF from the IMSLP[5]. The amount of transcription work was adjusted, based on the number of participants in each session.

Recruitment

Due to the COVID-19 crisis, all studies were conducted through online videoconferencing. For both rounds of studies, a similar protocol was followed. First study was conducted in 5 sessions, with 30 participants in total. The second study took place 4 months later and it was conducted in 4 sessions, with 33 participants in total. The participants of both studies were members of Dutch youth orchestras, namely: Collegium Musicum[6], Quadrivium[7], NJO[8], Sweelinck[9], Nijmeegs Studentenorkest CMC[10], Amsterdams Studenten Orkest[11], S. M. G. 'Sempre Crescendo'[12] and Almeers Youth Symphony Orchestra[13].

Focus Group Structure

Our study's goals were to gain insights on how and to what extent student orchestras use digital scores in their rehearsals and performances, but also the extent of their familiarity with digital tools for music transcription. To that end, we selected the Focus Group Discussions (FGD) methodology to collect qualitative data on the topics of our discussions, which would later help us understand how to better conduct online transcription campaigns. During the workshops, there were always two researchers present. One acted as the facilitator of the discussions and was aware of challenges that student orchestras might face, empathising better with the participants, while the second acted as the notetaker. Both co-curated the discussions and the FGD were conducted in English.

[4] https://www.krashna.nl/en/
[5] https://imslp.org/wiki/Sextet_in_E-flat_major%2C_Op.71_(Beethoven%2C_Ludwig_van)
[6] https://www.collegiummusicum.nl/en/
[7] https://www.esmgquadrivium.nl
[8] https://www.njo.nl/english/orchestra/orchestra
[9] https://www.sweelinckorkest.nl
[10] https://www.nijmeegsstudentenorkest.nl
[11] http://www.amsterdamsstudentenorkest.nl/en/
[12] https://www.smgsemprecrescendo.nl
[13] https://www.stichtingajso.nl/english/ajso/

Due to the diverse technical background of our participants in both parts of our study, we curated the discussion points in such a way to find how familiar they were with digital means to access and edit their selected music scores. This helped us to adjust the depth to which we discussed the technical aspects of our work.

Before the start of each workshop, we handed each participant a consent form, where they were informed about the discussion notes we would log and the option to record our session. We made explicit that any data will be treated confidentially and that if any participant would not feel comfortable to record the session, we would not do so. They were also free to withdraw their participation at any point, and they could choose to participate with their video camera on or off.

During the FGD, we demonstrated to the participants online interfaces, with which they were encouraged to interact and voice their opinions and discuss points to improve on their design. To quantitatively measure the users' perceived satisfaction, we distributed the Post-Study System Usability Questionnaire (PSSUQ) to all the participants. Answers to all the forms alongside all of the discussion notes, were treated anonymously without any identifying features.

Even though the study was not conducted in person, we believe that its effectiveness was not hindered. Participants were already familiar with online video calls at the time of the study and we were able to conduct several workshops per day with participants who would otherwise be more difficult to gather together in a physical meeting point. Finally, the systems and mockups used were all available online, making them easily accessible by all the participants through their device of choice.

Predefined discussion points For the purpose of our evaluation study, we conducted all workshops using the same outline of discussion points. As mentioned before, we adjusted the extent of technical details discussed, based on the background of the participants per workshop. The list below represents the main discussion points used as a guide during the workshops.

To assess the familiarity of the participants with digital music scores and discuss their use of them alongside digital transcription tools, we used the following questions:

- How familiar are you with semantically rich music scores?
- Do you use this kind of digital scores personally?
  If yes: When (e.g. during practice) and/or how (e.g. tablet)?
  If no: Why and what would make them more appealing?
- Do you use digital scores as an orchestra?
  If yes: How do you incorporate them?
  If no: Have you thought about using them? What are the reasons you haven't so far?
- How could/(or already do) you benefit from using digital scores?

Since Sciptoria system is focused on providing high-quality digital transcriptions of music scores, we wanted to find how familiar were the participants with Optical Music Recognition (OMR) workflows and tools. We discussed with them the following points, in order to showcase why it is a hard problem and what is the state-of-the-art.

- How we want to incorporate the crowd into OMR processes;

- Current state and limits of research.

We followed our discussions on OMR, with how the TROMPA project is working to improve OMR workflows by incorporating human-in-the-loop solutions. We discussed the transcription and improvement campaigns that we are planning and how they are organised. Finally, we explored the concept of crowdsourcing and how it can be applied on music score transcription workflows.

By the end of the FGD, we followed our usability tests on the provided mockups with discussions about the task mockups (opinions, ideas on how we can improve it etc). After this, participants had a better sense of what could happen within a campaign, and we continued to discuss motivational considerations to take into account, when seeking to run such campaigns as part of the practice of the participants:

- How such a campaign could benefit their orchestras;
- How could they see themselves motivated to participate in such a campaign;
- What about campaigns from other orchestras;
- How could the general public (non-experts) help such campaigns;
- Types of tasks they believe they could successfully do.
- What could be a satisfying final product coming from such a campaign?
- What music score elements could be tolerated to be missing?
- In case of less than 100% coverage, what extent of transcription coverage would be satisfying enough?

Usability tests As explained before in this chapter, we followed different usability tests on the workshop with Krashna Musika, which focused on testing the Campaign Manager and an existing transcription campaign; while during our workshops with other student orchestras and young professionals, we conducted usability tests on different Task designs. In all workshops, we measured the users' perceived satisfaction while using the provided interfaces, using the Post-Study System Usability Questionnaire (PSSUQ), distributed to all the participants. More specifically, the questionnaire contained the following statements:

1. Overall, I am satisfied with how easy it is to use this system.
2. I was able to complete the tasks and scenarios quickly using this system.
3. I felt comfortable using this system.
4. It was easy to learn to use this system.
5. The system gave error messages that clearly told me how to fix problems.
6. Whenever I made a mistake using the system, I could recover easily and quickly.
7. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
8. It was easy to find the information I needed.
9. The information was effective in helping me complete the tasks and scenarios.
10. The organisation of information on the system screens was clear.
11. The interface of this system was pleasant.
12. I liked using the interface of this system.
13. This system has all the functions and capabilities I expect it to have.
14. Overall, I am satisfied with this system.

Recruitment prospects Finally, we ended all the workshops by showcasing our plans for the TROMPA transcription campaigns and sought to motivate and encourage participants to also consider participating in future campaigns and user studies, and help us with recruiting further participants from their own circles, as soon as new studies will be conducted. Furthermore, as a token of gratitude for their time (with a workshop lasting 2h30 on average), participants were offered the choice between a membership to Entrée –the RCO's youth audience association– or an RCO CD.

3.1.4. Results and Improvements

First study

As expected, the musical background survey indicated that many of the players had extensive musical instrument training, representing a considerable diversity of instrument experience. While this could indicate different expertise on specific music notations (such as types of clefs), nevertheless, the UI and transcription tasks of those notations were designed to be primarily based on visual recognition of similar artifacts.

In all sessions, participants managed to fully complete all tasks. We found that the rhythm and pitch transcription tasks are much more time-consuming compared to the clef, key signature, and time signature tasks. The qualitative feedback by the participants indicated that the UI design of these tasks could still be further improved.

For the **time signature detection** task, participants suggested to include buttons for commonly occurring time signatures to further minimize the need for textual input. They also indicated that the **key signature detection** task was confusing for some, due to the high complexity of the annotation (key signatures can occur in multiple places, even in a small segments).

For the **rhythm transcription**, the UI was found cumbersome and it was suggested to expand to more buttons with common preset choices. Furthermore, the absence of note beams when transcribing, made the visual comparison between the reference and the entered input more difficult.

Regarding the **pitch transcription**, participants indicated that the task involved elaborate user input and that it would be useful to include shortcuts for common actions and input dragging. Furthermore, at the moment of these studies, the way the default pitch of the rhythm notes was registered during the previous task, was deemed insufficient to clearly visualise the notes in this task.

Execution time is estimated based on commit logs of the Git repository. Through the user evaluation we identified issues with this method for two main reasons: there may be time lag between subsequent commits due to the input volume processed by our system; results that did not alter the MEI snippet, where not committed (e.g. indicating no clef in given segment).

Finally, general feedback focused on the inconsistency of the 'submit' button's look and feel across tasks. Also, task instructions were found either unclear or too 'wordy'. These usability issues were also apparent in the PSSUQ survey results.

Implemented improvements

As noted in the previous section, there were some issues with the initial GUI designs that impacted user efficiency and the overall user experience. To rectify this, we implemented multiple changes.

For **time detection**, we added preset buttons with frequently occurring time signatures. For **key signature detection**, contributors can select the type of key signature, and click a button to increment the count, which will show a preview of the key signature. For **rhythm transcription**, a full redesign of the GUI was performed, replacing the slider with expanded preset buttons (see Figure 3.2). In addition, note navigation/deletion has been replaced by a single undo button. Finally, beam support has been added. For **pitch transcription**, octave adjustment buttons have been added, and notes get initialized in the middle of a staff, depending on the preceding clef. Furthermore, for note navigation, keyboard shortcuts were now implemented.

In terms of general improvements, the 'submit' button was standardized in terms of look, feel and location across tasks. Furthermore, the help text was replaced by a help button, launching a floating window, that shows an animation of how the task is supposed to be performed, along with a description. Finally, Verovio, the used score online editor, was no longer loaded for tasks that do not use any MEI preview, which improved loading times for the time, key signature and rhythm transcription tasks. During any interface loading, a loading progress indicator was also included. In the back-end, we also improved system logging, so more refined timing information could be included in our analyses (e.g. registering MEI snippet submissions with no alterations).

Second study

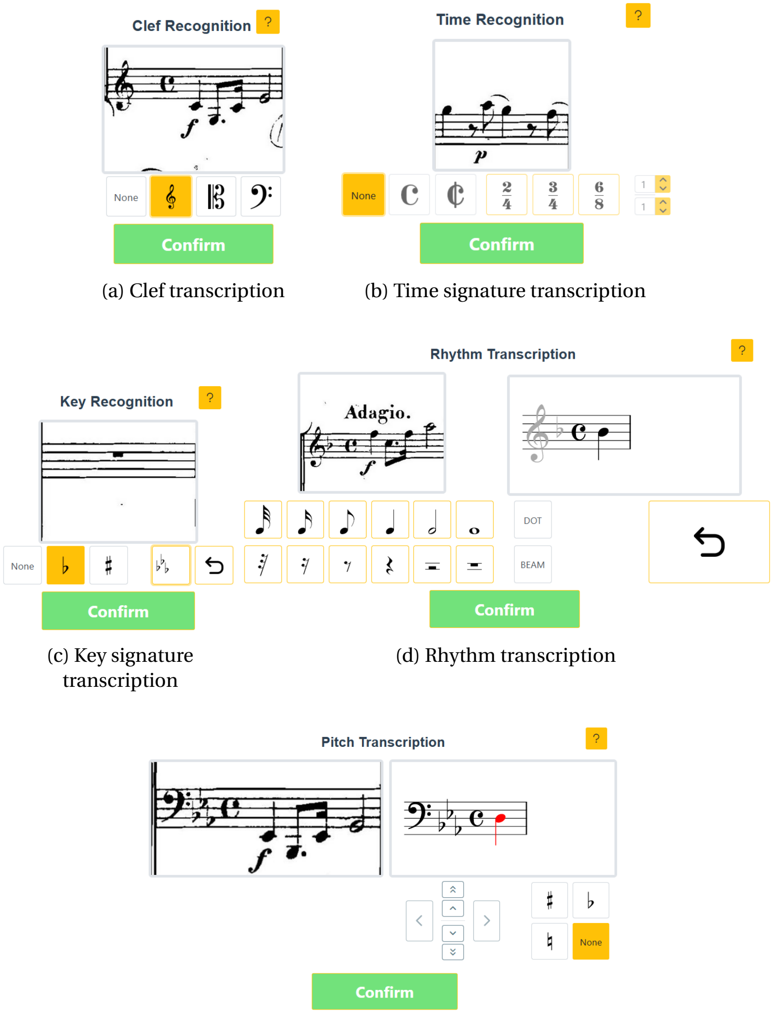
Following the results of the first study, the musical expertise of the participants was equally high and diverse in terms of instrument of choice. Following our improvements based on collected feedback, the participants in all sessions of the second study were very enthusiastic about our system. We identified major improvements in the amount of time spent per task, while also the both the feedback from discussions and the PSSUQ questions was positive.

Our improvements in the UI of **rhythm transcription** and **pitch transcription** tasks seemed to also assist better the users, to complete successfully their tasks.

3.1.5. Insights for Future Versions

From our discussions with the RCO experts and the members of youth orchestras, we received invaluable insights on the traditional transcription methods employed by professionals in classical orchestral music. We first-hand witnessed challenges, such as: messy handwritten annotations that obscure the printed music notations, imperfectly scanned pages and damaged scores. Although such challenges exist in all types of printed music scores, the length and complexity of orchestral pieces amplifies them, resulting in automatic transcription methods to frequently fail. Professionals and amateurs alike, still lean on manually transcribing scores from the ground up, using dedicated software, online solutions or even in cases, pen and paper.

Participants in our Focus Groups, discussed extensively their transcription habits. The majority of the youth orchestras relied on one or two people who transcribed the score for all the rest. Our microtask approach was happily welcomed and the participants indicated that a collaborative, task-based workflow could potentially improve productivity and the social bonding of a group.



(a) Clef transcription

(b) Time signature transcription

(c) Key signature transcription

(d) Rhythm transcription

(e) Pitch transcription

Figure 3.2: Improved designs for the transcription tasks.

Our discussions with the participants brought another valuable insight on the annotation needs of orchestras: almost unanimously the participants pointed towards sharing performance annotations between orchestras. When performing music pieces, each orchestra adds their own interpretation that can often be quite unique and separate from others. The potential of digitizing and sharing those performance annotations on top of the other music notations and sharing them between users of the transcription platform, was deemed to be a key future to its success.

3.2. Error Detection on Music Scores

This section contributes towards a better understanding of how music transcription could be supported, and potentially scaled up, through microtask crowdsourcing. A transcription system such as the one described in Section 3.1, can still be prone to errors introduced either by an automated process or by human error. To that end, we focus on a simple yet fundamental problem: the identification of differences (errors) between two music scores segments. Spotting errors is, in itself, a very useful operation in music transcription workflows, as it could be of assistance for experts transcribing a score, with the creation of labeled data to train automated OMR systems, or with the identification of errors made by such a system. Workers operating in online microtask crowdsourcing platforms are already accustomed to such type of tasks, but the understanding of music scores is not a common skill.

The main research question addressed in this work is:

- **[RQ5]** To what extent can workers from microtask crowdsourcing platforms detect errors in transcribed music scores?

To answer the question, we setup an experiment where two microtask design factors were adjusted respectively, the score transcription's *modality* (spotting errors on visual vs. audio), and the *size* (in terms of measures) to be analysed. We recruited 144 workers from Mechanical Turk, asked them to check 144 segments for several types of errors, and measured their performance in terms of completion time, accuracy, and sustained cognitive load.

Results show that crowd workers were able to achieve maximum precision of 94% and accuracy of 85% using an interface that combines visual and audio modalities, thus showing that microtask crowdsourcing is useful for error detection, and that workers benefit from the audio extract of the transcribed score, both alone and as a support for the visual comparison.

3.2.1. Related Work

The topic of microtask crowdsourcing for music transcription is scarcely addressed in literature, with many relevant research questions left unanswered. In Burghardt et al. [14] the *Allegro* system was developed, a tool to allow the transcription of entire scores by a (single) human worker. However, *Allegro* has only been tested on a limited number of users, and it was not deployed on an online microtask crowdsourcing platform. The same limitation holds for the work in [20], one of the first attempts to design a human-in-the-loop approach for OMR, studying how task design and their recognition engine can be optimally combined with human input. This study focused on analysing segments which are one measure long, which is the smallest unit of analysis in our study as well. We expand this, by studying also how the size of the segment shown to the crowd affect its performance.

An important work to mention is OpenScore [30], up to now the largest scale project to incorporate humans in music score transcription. In terms of user participation though, it was mainly carried out by seven community members with extensive musical background. Moreover they report different issues related to the management of data (done manually by the administrators of the platform) and user engagement (without any control they would focus on their preferred music score) admitting in the end that in their project "OMR (involving humans) is not currently a scalable solution".

So far, there is not any literature that has targeted unknown crowds with varying skills for music transcription tasks, thus research questions on [80] about what type of tasks users can perform and how to evaluate them still remain open. In this work we address this research gap by looking into similar crowdsourcing works in other domains. More specifically, in [69] it was found that for knowledge-intensive tasks involving artworks, a crowd with varying and unknown domain-specific knowledge found on online platforms can produce useful annotations when aided by good task design.

Research has shown that UI design is an important part of a microtask design [25]. Research so far has experimented with various designs such as showing spectogram visualisations for audio annotation [16] or the use of chat-bots to assist common types of microtasks [62], all of which have yielded positive results on the performance of the crowdworkers. Inspired by this we make the design of the work interface a central point of our study.

3.2.2. Experimental Design

The main focus of this work is to study to what extent a general crowd can identify *errors* in a music score transcription. We therefore designed an experiment aimed at testing the ability of crowd workers to spot errors using interfaces having a combination of visual and audio components.

Task Design

Our aim is to study how different task design factors can influence the crowdworkers performance, focusing on two aspects:

1. The *modality* (*visual* versus *audio*) used to spot errors: as music scores are complex artifacts, and music is primarily an auditory experience. Therefore, we investigate how the score comparison *modality* affect the error detection performance in workers that are potentially not familiar with musical notation. Intuitively, we want to investigate if "hearing" errors is easier that "seeing" errors.

2. The score *size* offered to crowdworkers for annotation. The goal is to assess how the size (in terms of measures) of the score offered to worker affects their performance.

To develop a better understanding on the characteristics of the crowd, we open the tasks with questions about their demographic information (occupational status, level of education and age) and their music sophistication. For the latter, we compiled a list of 6 questions from The Goldsmiths Musical Sophistication Index (Gold-MSI) [65].

Crowd workers' performance with error identification is measured using accuracy, precision and recall and time to execute a microtask. In addition, we measure user confidence with their judgements with a seven-value Likert scale.

Finally, we measure the sustained cognitive load when executing the microtask, measured through the NASA Task Load Index (TLX)[14], which ranges from 0 to 100 (higher the TLX is, the heavier cognitive load the worker perceives).

[14] https://humansystems.arc.nasa.gov/groups/TLX/

Evaluation Dataset

Selecting a suitable music score was our first step preparing our experiments. We use a single classical music score to avoid introducing additional variable in workers' performance. Specifically we use the Urtext of "*32 Variations in C minor*" by Ludwig van Beethoven. It is a piano piece and the music artifacts are all printed typeset forms. This is a slightly easier use case than hand-written scores. The score was retrieved from IMSLP as a PDF[15].

As a Gold standard transcription of that PDF we used an MEI[16] file that had been transcribed by an expert. This file was accepted as error free, and it allowed us introduce errors in a controlled way for our experiments.

We segmented the music score in varying sizes to investigate how workers cope with shorter or longer tasks. We distinguish 1) *one measure* segments, 2) segments of *two measures* and 3) segments of *three measures*. Both of the two digital versions of the score, the PDF file of the original score and the transcribed MEI file, were segmented using the aforementioned segment sizes. The segmentation of the PDF was performed manually, while for the MEI we used the appropriate identifiers of each measure that was included in the corresponding image segment, to isolate the correct headers in the MEI. Since it's a piano score, each measure contains two staves.

We then introduced errors in the MEI segments derived from common errors that can occur in automatic OMR systems. The type of errors could impact the crowdworkers' ability to spot them and correctly identify them as errors. Some of them can be challenging to notice even to experts of the field. Therefore, we study different types of errors, all focusing on the music notes themselves and their accidentals. Errors on performance annotations, clefs, finger numbers etc, are out of scope in this study. We introduced the following types of error per MEI segment: 1) *Missing notes*; 2) *Wrong vertical position of a note*; 3) *Wrong duration of a note*; 4) *Wrong accidental*.

Each segment that was shown to the user contained only one type of error. The amount of errors per segment was kept constant at 40% of the notes present in the segment. Thus, if a crowdworker is presented with two measures with notes missing, then notes are missing on both measures at a 40% rate of the total notes on both measures combined. No more types of error are present in the segment.

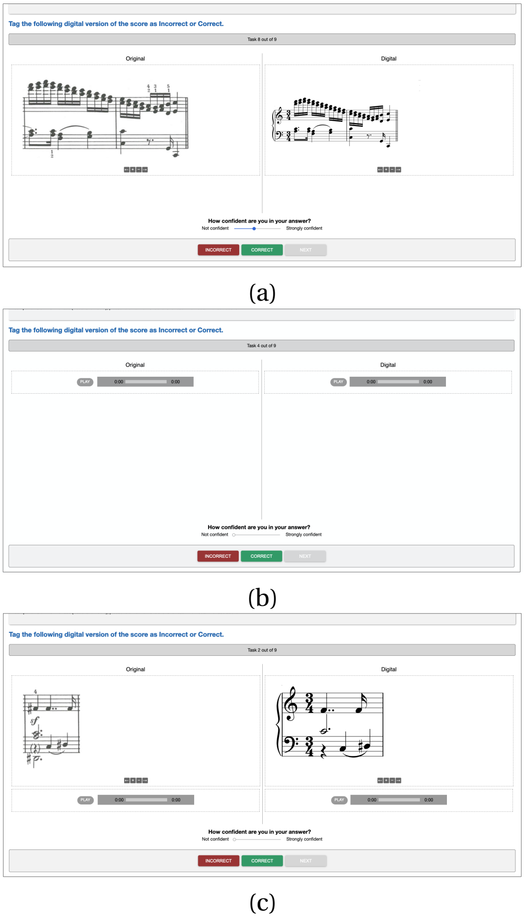
To make the performance comparisons meaningful, we ensured that our dataset is balanced across all error types. In total we used 144 segments derived from the entirety of the selected piano score, with 48 segments per size category, from which 24 were equally distributed to each type of error, while the remaining 24 were kept correct.

User Interface Design

To test the modalities' effects separately and accurately, we designed three different interfaces: one that would have image-to-image comparison to test the traditional form of the task, one with only audio-to-audio comparison, and one with both audio and image comparison. The interfaces are designed to include the following data. 1) **Original Score**: the segment's image from the scanned score. 2) **Correct MEI Render**: a render of the transcribed version of the *Original Score*'s segment; 3) **Incorrect MEI Render**: a

[15] https://imslp.org/wiki/32_Variations_in_C_minor%2C_WoO_80_(Beethoven%2C_Ludwig_van)

[16] https://music-encoding.org/


Tag the following digital version of the score as Incorrect or Correct.

Task 8 out of 9

How confident are you in your answer?

(a)

Tag the following digital version of the score as Incorrect or Correct.

Task 4 out of 9

How confident are you in your answer?

(b)

Tag the following digital version of the score as Incorrect or Correct.

Task 2 out of 9

How confident are you in your answer?

(c)

Figure 3.3: Microtask User Interfaces: (a) Visual, (b) Audio and (c) Combination

render of the MEI transcription containing errors. 4) **Correct MIDI**: the MIDI extract of the correct version of *MEI Transcript*'s segment. 5) **Incorrect MIDI**: the MIDI extract of *MEI Transcript*'s segment containing errors.

We refrained from using audio from a recorded performance against a MIDI extract containing errors to avoid confusing the crowd on what constitutes as “different” or an “error” in the audio comparison task. A recorded performance would introduce performance-related artifacts to the audio, which do not exist in a MIDI extract, thus increasing the chance of false negatives in identifying an audio snippet as "incorrect". Finally, for the combined comparison, we used the same elements as with the individual comparisons. For the renders of the MEI transcripts and MIDI extracts we used Verovio[17] on our interfaces.

From a design perspective, the interfaces needed to be simple and closely resembling

[17] https://www.verovio.org/index.xhtml

each other to minimize their effect on the workers' judgements. They should also be able to facilitate the different segment sizes without changing the layout. Eventually, we also wanted to accommodate error detection in the same manner for both image and audio comparisons, to avoid differences on the annotation tools being another factor to the crowd's performance. Therefore, we designed the error detection task to ask from the users to annotate a given MEI transcript or MIDI extract as "*Incorrect*" if it exhibit errors and as "*Correct*" if they did not.

In all interfaces, the segments to the left are associated to the original scanned score and the correct MEI transcription of it, while to the right we always place the segments that need to be annotated. The MEI render or the MIDI extract to the right can be either "*Correct*" or "*Incorrect*" and we calculate the performance of the workers based on identifying this correctly. In addition to the two buttons for the labels, we included a slider to indicate the worker's confidence in their label. Later in the analysis of the results, we used this indicator to study how each interface and segment size affected the confidence of the workers' to their judgements. These design considerations resulted in the following three interface designs:

- **Original Score** against **Correct / Incorrect MEI Render (Visual)**: This user interface, depicted in Figure 3.3(a), shows the segment of the original scanned score to the left, with the corresponding MEI render to the right. The user needs to compare the two images and spot differences related to the types of errors.
- **Correct MIDI** against **Correct/Incorrect MIDI (Audio)**: In this interface, as shown in Figure 3.3(b), we let the user listen to the correct MIDI extract on the left and the one generated from the MEI transcription to the right.
- **Original Score and Correct MIDI** against **Correct / Incorrect MEI** and **Correct / Incorrect MIDI (Combination)**: This final user interface, as shown in Figure 3.3(c), combines elements of the previous two. The user here has the option to either use the visual comparison, the audio comparison or both to realise if there are errors to the segment to the right. The MEI render and MIDI extraction to the right always originate from the same MEI transcription, therefore both will be correct or both will contain errors.

Each combination of interface with a segment size consists of a microtask. To efficiently and effectively gather performance data, we wanted the same worker to be "exposed" to all nine possible combinations. Therefore, in its final form, a worker would have to execute a task that would begin with a set of demographic and music sophistication questions, followed by the nine microtasks and end with the cognitive load questionnaire. To minimise the impact of issues related to the familiarity of workers with the interface, the task also includes an introductory explanation of the work interface, with examples of errors and expected responses. The results of our experiment are analysed based on the overall, but also on error type, performance of workers.

3.2.3. Results

As discussed in previous sections, we published our tasks on Amazon Mechanical Turk (MTurk). The platform offers several configurations for each batch of tasks submitted. We published them as public so they can be accessed by all the users of the platform and we

Interface	**segment size**		
	1	2	3
Visual	**P=66.22** R=59.76 A=60.27	**P=65.28** R=61.84 A=62.24	P=59.72 R=74.14 A=69.23
Audio	P=60.27 **R=81.48 A=72.92**	P=62.50 **R=81.82 A=74.13**	P=64.79 R=80.70 A=74.83
Combined	P=59.70 R=68.97 A=67.63	**P=65.28** R=74.60 A=71.33	**P=68.06 R=83.05 A=76.92**

Table 3.1: Precision, Recall and Accuracy of *individual* answers, by segment sizes and interfaces.

did not require any Mechanical Turk Master (expert workers). Only to avoid malicious workers, we filter them by their previous HIT Approval Rate, and we set it to 95%.

In total, 144 workers executed our tasks on MTurk and we paid them per task execution according to the average US minimal hourly wage[18]. In order to minimize the effect of any biases or learning effect we randomized the order of the presentation of the different task designs (UI-segment size combination). One worker eluded the quality verification on task interface, which results in 143 unique workers.

Worker Demographics

From a demographic aspect, most of the workers (84.6%) reported that they had a full time occupation. Also, 67.8% of total workers reported their education level was Bachelor's degree, while 14.9% of them had Master's degree. Only 8.3% of the workers were above 45 years old.

Answers to the Gold-MSI questions indicate that the majority of workers seem to be familiar with listening to music, as 56% of them listen to music for at least 1 hour a day and 65% say they can hit the proper notes while listening to a record. Also, the majority of them (75%) state they can properly compare and discuss different performances of the same music piece. On the other hand, 52.4% of the workers reported having up to one year formal training in music theory, where the 26.6% has no prior education on the subject. Also, 41.95% of the workers have trained for maximum one year on a music instrument, while 22.4% of never had. Their answers here suggest that the majority of the crowd has little to no music theory background, and a considerable amount of them also no formal studies on an instrument. The results also suggest that the crowd executing our tasks was mainly composed of workers with little expertise with music theory.

Interface	**segment size**		
	1	2	3
Visual	P=60.71 **R=70.83** A=62.50	P=67.86 **R=79.17** A=70.83	P=88.89 R=66.67 A=79.17
Audio	**P=100.0** R=62.50 **A=81.25**	**P=88.24** R=62.50 A=77.08	P=90.00 R=75.00 A=83.33
Combined	P=76.47 R=56.52 A=70.21	P=85.00 R=70.83 **A=79.17**	**P=94.74 R=75.00 A=85.42**

Table 3.2: Overall Precision, Recall and Accuracy of *aggregated* answers by segment sizes and interfaces. In bold you find the highest precision, recall and accuracy by segment size, while underlined you find the highest overall performance

[18] We estimated an average task completion time of 15'; each crowdworker was awarded 2.5$

ACCURACY

The results per target segment were aggregated from three different individual workers. Table 3.1 shows that tasks performed with the *Audio* interface consistently achieved higher accuracy than the *Visual* one. The *Combined* interface achieved good accuracy figures with all segments sizes, but best accuracy with the 3-measure-long segments. The *Visual* interface yielded consistently the lowest recall and accuracy results, for all segment sizes. Interestingly, the precision for this interface on segment size one, was the highest compared to *Audio* and *Combined* for the same size.

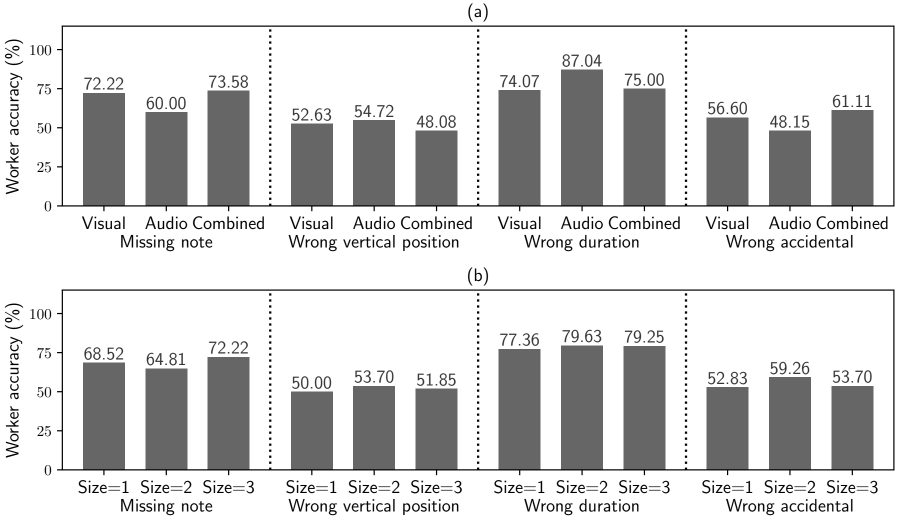


Figure 3.4: Workers error detection accuracy (unit:%) (a) per user interface and (b) per segment size.

Figure 3.4 shows the accuracy of the workers in detecting specific type of errors. *Wrong duration* error seems to be accurately spotted in any user interface and segment size, with the *Audio* interface resulting in the highest accuracy (87.04%). Workers perform better detecting the *Missing Note* error using the *Combined* interface and the 3-measure segments. The accuracy obtained with the *Visual* interface though, suggests that workers might rely more on the image of the score rather than the audio for this type of error. The *Wrong Vertical position* error was more difficult to detect in general; the highest accuracy was obtained with the *Audio* interface (54.72%) and with the segment size of 2 measures (53.70%). Finally, the *Wrong accidental* type was the second most demanding error to be detected with the highest accuracy achieved using *Combined* interface (61.11%), with a slight improvement in segments containing 2 measures.

In microtask crowdsourcing it is common to aggregate individual results to improve overall quality. Table 3.2 shows the performance achieved using a simple *majority voting* aggregation scheme. The *Combined* interface with 3-measure-long segments still achieves best performance with a remarkable 94% in precision, and 85% in accuracy. The *Audio* interface achieves best precision performance for both 1-measure-long and 2-measure-long segments, while the *Visual* interface achieves best recall.

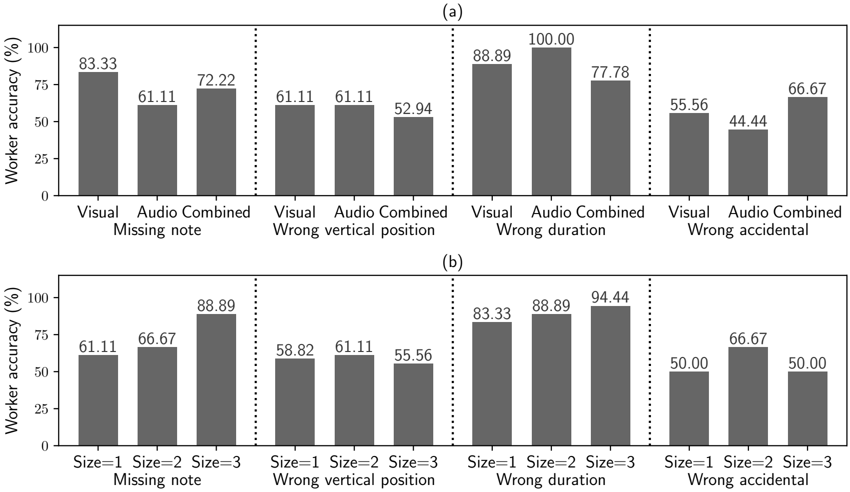


Figure 3.5: Aggregated error detection accuracy (unit:%) (a) per user interface and (b) per segment size.

Figure 3.5 shows the aggregated accuracy in detecting specific type of errors. In terms of *Wrong Duration* error, the accuracy remains the highest after the aggregation. The *Audio* interface and the 3-measure-long segments achieve 100% and 94% in accuracy respectively. *Visual* interface and the 3-measure-long segments obtain the highest accuracy (82% and 88% respectively) in detecting *Missing note* error. The *Wrong Vertical position* error and the *Wrong accidental* error still have relatively low accuracy.

EXECUTION TIME

Figure 3.6 shows that, as expected, execution time generally increases as the segment size increase.We performed statistical tests (independent t-tests, $\alpha = 0.05$) to find significant differences between interfaces and segment sizes. In the case of *Wrong vertical position* error though, the *Audio* interface allowed the worker to spot the errors significantly faster compared to *Combined* interface ($p = 3.5\text{e-}3$). For the *Wrong duration* error the addition of audio and the increased segment size can lead to a significantly longer average execution time (for both *Audio* and *Combined* interfaces compared to *Visual*, $p = 1.3\text{e-}4$ and $p = 1.9\text{e-}5$ respectively; and for 3-measure-long segment vs 1-measure-long segment, $p = 2.5\text{e-}3$).

For the *Wrong accidental* case, we see that worker spent less execution time on the *Visual* interface (no significance). However, comparing it with the results, the worker most probably dismissed the segment as "Correct", rather spend more time in case they had missed the error.

MUSIC SOPHISTICATION AND COGNITIVE LOAD

The average score of cognitive task load (NASA-TLX) is 47.7%, a typical value for classification and similar cognitive tasks [31]. To investigate how music sophistication relates to worker performance and cognitive load on spotting music errors, we select and analyze

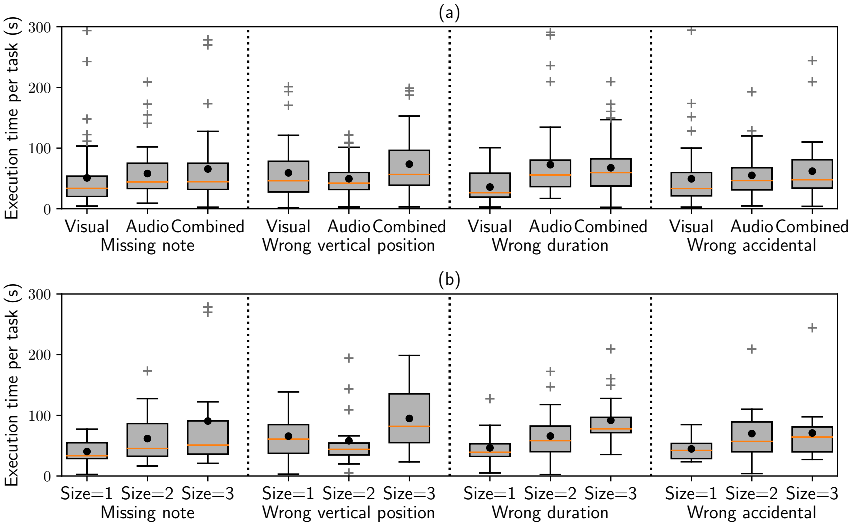


Figure 3.6: Worker execution time (unit: seconds) of each microtask by (a) user interface and (b) segment size.

a corresponding question from Gold-MSI questionnaire – “I find it difficult to spot mistakes in a performance of a song even if I know the tune.”, where workers need to select an option from `Completely Disagree` to `Completely Agree`, before they execute the music transcription tasks.

Results show that 47% workers agreed with the statement that it was difficult to spot mistakes in a performance of a song; 33% of them disagreed with the statement, and the rest of them (20%) were unsure. We calculated the worker accuracy and the cognitive task load score, and performed significant testings (independent t-test, $\alpha = 0.05$). Workers who reported lower difficulty with spotting music errors (accuracy = 81 ± 15%, TLX score = 44.36 ± 13.05) outperformed workers who had higher difficulty (accuracy = 63 ± 16%, TLX score = 50.86 ± 13.61) in terms of both worker accuracy and perceived cognitive load (p = 3.3e-8 and 0.013 respectively). The workers who reported lower difficulty also had significantly higher accuracy (p = 0.011) compared to unsure workers (accuracy = 0.70 ± 0.20%, TLX score = 46.01 ± 14.27). Results suggest that the self-reported music sophistication in some specific aspects strongly relates to actual worker performance in error identification and cognitive load. Nonetheless, workers with lower sophistication still achieved good performance, with a small additional cost in terms of cognitive load.

3.2.4. DISCUSSION

As expected, people with some formal knowledge in music, which could be useful to comprehend music scores, are very rare “in the wild”. To enable the use of microtask crowdsourcing for music score transcription, good task design is therefore of essence. Results show that error detection is a task that could be successfully performed in a microtask crowdsourcing setting. Offering audio extracts of a target music score can positively affect the performance of the crowdworkers, especially for short segments of one or two measures. With larger segments, even though audio extracts are still yielding better results against to the textual measures of the score, a combination of the two modalities is more preferable. This result gives important indications for task splitting and scheduling purposes, as it suggests that it is possible to evaluate larger portions of scores without incurring accuracy penalties. This has obvious implications in terms of overall transcription costs.

In terms of types of detected errors, results suggest that the *Missing Note* and *Wrong Duration* errors are the easiest to be found, while the crowd had relatively more difficulty detecting *Wrong Accidentals* and *Wrong vertical position* ones. Furthermore, we see the clear effect of user interface and segment sizes in identifying correctly specific errors. Specifically, the *Audio* interface helps in finding *Wrong duration* errors, while the *Combined* one increases the accuracy in finding *Wrong accidental* mistakes. Showing segments longer than two measures seems to slightly hinder the ability of the crowd to detect any errors besides *Missing notes.*

Limitations. Correct MEI render and correct MIDI files of scores to be transcribed are typically not available in the real world. The distribution of errors in the evaluation dataset might not reflect the actual distribution of errors produced, for instance, by OMR systems. Given these limitations, the results of our experiment are probably to be interpreted as an "upper bound" in terms of achievable performance; nonetheless, they clearly indicate that the detection of errors in transcribed music score is an activity that can be successfully performed by crowdworkers.

3.2.5. CONCLUSION

Music score transcription is an important activity for written music preservation. Through this work, we show that microtask crowdsourcing can be used to scale up specific transcription activities. Worker interfaces that combine visual and audio modalities allow the evaluation of longer score segments. Focusing on the error detection task, results show that crowd workers can achieve high precision and recall, especially with *Missing Note* and *Wrong Duration errors.* In future work, we plan to expand the evaluation dataset, perform experiments where workers are asked to compare recorded performance, and address a broader set of transcription errors. Finally, we will investigate other types of microtasks, and study to what extent crowd workers could also be employed to *transcribe* scores.

3.3. TEMPORAL ACTIVITY DETECTION OF MUSICAL INSTRUMENTS IN AUDIO

Musical instrument recognition enables applications such as instrument-based music search and audio manipulation, which are highly sought-after processes in everyday music consumption and production. Despite continuous signs of progress, advances in automatic musical instrument recognition are hindered by the lack of large, diverse and publicly available annotated datasets. In the case of classical music, where symphonies are composed with a plethora of different instruments in mind, this is especially true.

Instrument recognition in classical music could tremendously help preserve information from live orchestrations, capturing and showcasing the changes that conductors have made to compositions. This can be particularly important for cases where the original handwritten annotations and changes haven't survived the test of time (especially considering the fragility of paper as a medium), or when the scores are forgotten in private collections. Therefore, capturing this information through audio recordings will help preserve valuable information on how different musicians have interpreted classical music compositions. Such a task though is particularly challenging, especially since different instruments might have similar texture and tone.

Crowdsourcing has been successfully utilized to scale annotation processes to meet the ever-higher demands of data-driven algorithms [67, 110, 89]. While works such as [37] and [92] show that crowdsourcing can be a viable and powerful tool to distinguish and annotate music audio, it still remains underutilised as a tool in the domain, primarily due to the complexity of the annotation tasks [32] which are believed to demand extensive domain knowledge and training – arguably, musical elements such as tempo, chords and timbre can be demanding for an untrained human annotator to detect.

With this study, we aim to provide more evidence that complex music audio annotation tasks can be performed on crowdsourcing platforms. We focus on the task of musical instrument activity detection and investigate non-experts' capability to recognise their activity and annotate the times in which they perform. Our study builds upon the findings of [37] where users were able to detect if an instrument was present in an audio excerpt or not. We extend this detection task to also cover the exact time-frames of instrument activity. This is a type of task where experts are commonly employed [11] to annotate data, due to several challenges such as multiple instruments playing simultaneously [33, 44], or instruments of the same family exhibiting similar timbre [109, 75].

More specifically, we explore and analyse the capabilities of crowd workers to effectively detect temporal aspects of musical instrument activity in polyphonic audio (with focus on trio ensembles). We seek to answer the following questions:

- **[RQ6]** To what extent can non-experts detect the onset and offset of a musical instrument's activity on polyphonic audio?
- **[RQ7]** How do their self-assessed perceptual abilities and musical knowledge relate to their performance?

Our study takes place on Prolific[19]. The audio excerpts were chosen from three different genres (namely *classical*, *jazz* and *rock*) to understand if different instruments and rhythms can affect the performance of crowd workers. We also utilize a set of pre-established and evaluated questionnaires to retrieve user attributes, that can potentially relate to their performance. We employ the "Musical Training" and "Perceptual Abilities" categories from Goldsmith's Music Sophistication Index (GMSI)[65], a questionnaire specifically designed to capture an individual's ability to engage with music. These specific categories were found previously to most significantly predict the workers' musical perceptual abilities[86].

Our results show that non-experts can demonstrate good perception of musical instruments' temporal activity for the chosen audio excerpts. Their self-assessed perceptual abilities reflect reasonably well their actual perception skill. These results open possibili-

[19] `https://www.prolific.co`

ties for further future studies on instrument activity annotation and provide a positive outlook for systems relying on such annotations.

3.3.1. RELATED WORK

The work in OpenMic 2018 [37] is one of the first attempts to annotate instrument presence for instrument recognition at scale, employing 2,500 unique annotators from CrowdFlower [20], using excerpts from Free Music Archive[21] and the AudioSet[28]. The researchers followed specific task design approaches to assist the crowd workers in their task, which they adapted after an initial study. The annotation process was limited to binary annotations, indicating the presence or absence of a musical instrument in an audio excerpt. Showcasing that crowd workers are able to provide strongly-labeled data, e.g. with temporal annotation, as in our study, can enable new opportunities for instrument activity detection and source separation.

Even though the study in Cartwright et al. [16] is not based on music audio, it nevertheless demonstrates the crowd's ability to annotate temporal aspects of audio events in audio recordings, a feat very relevant to our study. Our interface design is inspired by this study, as the crowd workers had to draw bounding boxes on spectrogram visualisations of audio excerpts. The sounds in [16] were synthesized using Scaper [81], for greater control over *max-polyphony* and *gini-polyphony* (amount of sound overlap).

Our study is also motivated by recent findings regarding crowd workers music perception abilities [86]. Users of crowdsourcing platforms were shown to possess considerable skills in detecting music aspects such as tempo and melody.

To the best of our knowledge, the current literature lacks works that study the performance of crowd workers on temporal activity detection of musical instruments in relationship with worker demographics or musical properties, which is the goal of this work.

3.3.2. EXPERIMENTAL DESIGN

We designed our experiment to study and understand if users on crowdsourcing platforms can perceive the temporal activity of a musical instrument in audio excerpts. We aim to focus on realistic use cases, thus testing the workers' capacity to perceive instruments in audio excerpts that are performed, recorded, mixed and mastered professionally.

Therefore, we used existing recordings instead of synthesized audio which would have been less representative of real-life scenarios, but could have given us higher control on the musical aspects of the audio and instrumentation. To that end, we carefully selected the audio excerpts to control, as much as possible, musical aspects such as timbre and performance.

We employed previously established and evaluated questionnaires, to learn about workers' (a) "Perceptual Abilities" and "Music Training" through Goldsmith's Musical Sophistication Questionnaire (GMSI); (b) cognitive load through NASA's Task Load IndeX (NASA-TLX) survey[22]; (c) equipment quality [77] and (d) outside noise [9].

[20] https://visit.figure-eight.com/People-Powered-Data-Enrichment_T

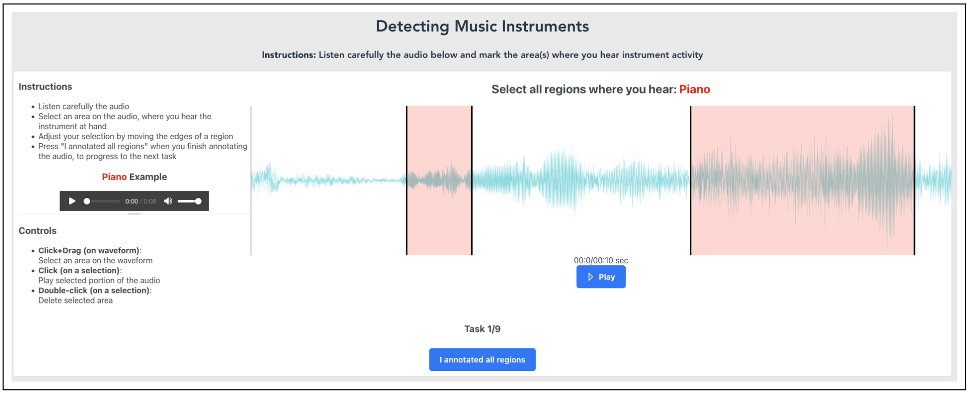
[21] https://freemusicarchive.org

[22] https://humansystems.arc.nasa.gov/groups/tlx/

The task workflow started with simple demographic questions, followed by the GMSI questionnaire. The user was then introduced with the main task to annotate audio excerpts. The study concluded with a post-task survey regarding their cognitive load, equipment and a general feedback entry.

Selected Audio Excerpts. For the main annotation task, we made use of audio excerpts from trio ensembles of three major genres, *classical*, *jazz* and *rock*. We used audio excerpts of these particular three genres due to their wide discrepancy in instrumentation and rhythm. Even though on some occasions the instruments used in each genre can showcase timbre similarities (like double bass and bass guitar), in other cases, the timbre can differ wildly (electric guitar compared to cello). To the best of our knowledge, there is no previous baseline of the crowd workers' perception of polyphonic music, so we decided to control for the maximum number of instruments that would play simultaneously in an excerpt, by selecting recordings of trio ensembles for each genre. Each audio excerpt had a length of 10 seconds, as used also in similar studies [37, 16]. The authors annotated the instrument activity per audio excerpt, which was later used to evaluate the crowd's annotations.

For the classical music excerpts, we made use of a specific type of a trio ensemble, namely *piano*, *clarinet* and *cello*. On the selected music clip, we selected an excerpt where both *clarinet* and *cello* have prominent parts, while *piano* is mostly following in the background. For our jazz excerpt, we used of the more standardized trio ensemble of *piano*, *double bass* and *drums*, where *double bass* and *drums* keep the rhythm and *piano* is performed in small melodic bursts. Lastly, for the category of rock, we made use of a music excerpt from "power trio" bands, which most frequently consist of *electric guitar*, *bass guitar* and *drums*. It follows the same performance pattern as the jazz excerpt on the *bass guitar* and *drums*, while the *electric guitar* enters near the middle of the excerpt with a sustained, distorted power chord.



Detecting Music Instruments

Instructions: Listen carefully the audio below and mark the area(s) where you hear instrument activity

Instructions

- Listen carefully the audio
- Select an area on the audio, where you hear the instrument at hand
- Adjust your selection by moving the edges of a region
- Press "I annotated all regions" when you finish annotating the audio, to progress to the next task

Piano Example

Controls

- **Click+Drag (on waveform)**: Select an area on the waveform
- **Click (on a selection)**: Play selected portion of the audio
- **Double-click (on a selection)**: Delete selected area

Select all regions where you hear: Piano

00:0/00:10 sec

Task 1/9

Figure 3.7: Main audio annotation task

We hypothesise that bass instruments will be more difficult to annotate in these genres, as bass-related sounds are more often "pushed back" during the mixing stage for such types of music. The different genres were selected to lessen the impact of possible enculturation bias. We believe that if only one genre was selected, participants who would be more familiar with it would find it easier to spot the activity of instruments prominent in the genre. With the selected genres, we cover a variety of rhythms, instrumentations and performative aspects, which could impose a challenge to non-experts.

Task and Interface Design. To assess the music expertise of the crowd we employed parts of GMSI, namely: "Music Training" and "Perceptual Abilities". The choice of the categories was based on a study on music perception skills of crowd workers [86], where results in these two categories were found to most significantly predict their auditory capabilities.

The questions of both GMSI categories were aggregated into one questionnaire, with one attention question placed in between the questionnaire's items. The users also had the ability to use a "Back" button to return to a previous question and alter their answer. We used the complete set of questions on both "Music Training" and "Perceptual Abilities", after consulting the online GSMI "configurator"[23].

The users were greeted with an "Instructions" message before the main annotation task, which described the steps to complete each microtask and a warning regarding the volume (as seen in Figure 3.8). The main audio annotation task (see Figure 3.7) consisted of four main parts: (a) audio waveform and controls (centre-right), (b) instructions and instrument example (upper left), (c) description of controls and (d) submission button with a simple progress indication. The instrument to be identified was indicated on both (a) and (b) in red, to draw the attention of the users.

Based on the findings during the OpenMic 2018 work[37], the crowd workers were found to struggle to detect multiple instruments at once. To that end, we followed their task design of annotating one instrument at a time; we presented the participants with the audio excerpt and requested to annotate the regions where a chosen single instrument, was active during the recording.

The worker would be presented with an audio excerpt and was instructed to detect the activity of one of the instruments present in the excerpt. The same procedure would follow for each of the instruments per audio excerpt, presented in a random order across genres (e.g. piano from classical music excerpt, followed by the electric guitar from rock music excerpt).

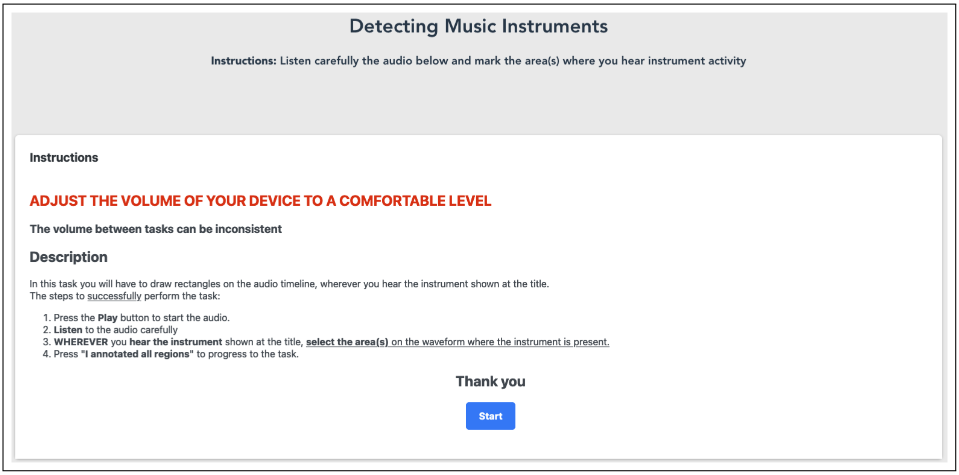
In the audio annotation interface, the users could play and pause the audio excerpt while also drawing bounding boxes on the audio waveform. The regions drawn on the waveform were adjustable on both ends and the user could easily dismiss them with a double-click. A single click on a region would play only the selected part of the audio excerpt. A crowd worker could only progress to the next excerpt if they had drawn at least one bounding box on the waveform.

For the design of the interface, we utilized `wavesurfer.js`[24] to draw the waveform and used the `regions` package to enable the bounding boxes interaction. Our choice of these tools was based on previous studies on audio annotation that utilized them successfully [16, 61].

Finally, as mentioned in [37], crowd workers could experience high cognitive load during instrument detection tasks, ultimately affecting their psyche. It was important for us to capture such a phenomenon, so we included the NASA-TLX questionnaire and a free text input to accommodate their feedback towards the study.

[23] `https://shiny.gold-msi.org/gmsiconfigurator/`

[24] `https://wavesurfer-js.org`



Detecting Music Instruments

Instructions: Listen carefully the audio below and mark the area(s) where you hear instrument activity

Instructions

ADJUST THE VOLUME OF YOUR DEVICE TO A COMFORTABLE LEVEL

The volume between tasks can be inconsistent

Description

In this task you will have to draw rectangles on the audio timeline, wherever you hear the instrument shown at the title.
The steps to successfully perform the task:

1. Press the **Play** button to start the audio.
2. **Listen** to the audio carefully
3. **WHEREVER** you **hear the instrument** shown at the title, **select the area(s)** on the waveform where the instrument is present.
4. Press **"I annotated all regions"** to progress to the task.

Thank you

Figure 3.8: Task instructions and warning

Evaluation methods. Our task design is based on one audio excerpt per genre (10 seconds), where a maximum of three instruments can playing simultaneously. As described before, per task, a worker had to draw the regions where they detect the activity of the selected musical instrument.

To evaluate their performance, we followed the same methods established in [63, 16] and in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [100]. We segmented each excerpt into 100ms-long frames which had binary values, depending on the presence or absence of the selected instrument. A frame is considered active when there is an overlap between the annotation region and any portion of the time interval of the frame. We believe that the frame's resolution of 100ms can help us to adequately assess the extent of crowd workers' precision when annotating the temporal activity of an instrument. Based on the ground truth values, we later calculated *Accuracy*, *Precision* and *Recall* of the workers' annotations. To evaluate their performance, we followed the same methods established in [63, 16] and in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [100]. We segmented each excerpt ($N = 3$) into 100ms-long frames which had binary values, depending on the presence or absence of the selected instrument. A frame is considered active when there is an overlap between the annotation region and any portion of the time interval of the frame. We believe that the frame's resolution of 100ms can help us to adequately assess the extent of crowd workers' capabilities to detect the temporal activity of an instrument. Based on the ground truth values, we later calculated *Accuracy*, *Precision* and *Recall* of the workers' annotations.

3.3.3. RESULTS

The study took place on Prolific, employing 30 crowd workers. We used the built-in prescreening filters of Prolific, setting criteria for fluency in English – for instructions' comprehension and higher chance of affinity to Western music – and minimum task approval rate to 90% – to maximise the chances for good-quality work. The reward was set to 4.5 GBP (5.62 USD) which was classified as "*Good*" by the platform. We preserved the results of the 28 workers (see their demographics in Table 3.3) that successfully passed the attention question. Filtering the results based on the attention questions.

	Variables	**Statistics**
Gender, n	Female	10
	Male	17
	Prefer not to say	1
Age (years)	Range	18-55
Occupation	Full-time	12
	Part-time	5
	Unemployed	11
Education	Associate degree	2
	Bachelor's degree	12
	High school/HED	4
	Master's degree	4
	Some college, no diploma	3
	Technical/trade/vocational training	3

Table 3.3: Participant demographics

Demographics and Equipment. The workers used mostly earphones, headphones and laptop speakers, while three reported using dedicated speakers. Most workers (15) reported the quality of their equipment as "Excellent", with the majority (22) reporting "Imperceptible" impairment. Finally, the majority (15) reported that conducted the study in near-silence conditions, while one reported performing the tasks in an environment with high noise levels.

Detecting Musical Instruments. The crowd workers showed high performance detecting most instrument activities on all three audio clips (RQ1). Studying the results per genre, we see in Table 3.4 that "Clarinet" was the most easily identifiable instrument. In the given audio excerpt, "Clarinet" had a prominent and distinct timbre, compared to the rest of the instruments. This might have helped annotators to detect its activity correctly. "Piano" on the other hand was more difficult to detect its temporal activity, as it accompanied the rest of the instruments with a softer tone.

	Accuracy	Precision	Recall
Piano	70.6%	91.5%	66.5%
Clarinet	**84.5%**	**95.8%**	**82.9%**
Cello	62.6%	95.5%	59.6%

Table 3.4: Accuracy, Precision, Recall and F-score on Classical audio excerpt (the highest scores per metric are in bold)

"Cello" though appears to be the hardest instrument to detect in the audio excerpt, as both accuracy and recall are near 60%. The high precision combined with low accuracy, could indicate that most workers mistook the activity of another instrument, with that of a cello. The results are surprising, as “Cello” was equally prominent as the “Clarinet”, playing at a lower register than the rest of the instruments.

In the case of “Jazz” we find the “Drums” to be the most recognizable instrument, while “Double Bass” yielded better results than “Cello” in the “Classical” excerpt (see Table 3.5. Recordings of “Double Bass” in jazz can vary from barely noticeable to accentuated, depending on the recording setting or the part of the song (being more prominent during solo performance). Despite being the prominent instrument alongside “Drums” for a large portion of the excerpt, the workers still had trouble identifying the regions where it was active.

	Accuracy	Precision	Recall
Piano	81.8%	70.9%	**87.7%**
Double Bass	64%	**100%**	64%
Drums	**84.4%**	**100%**	84.4%

Table 3.5: Accuracy, Precision, Recall and F-score on Jazz audio excerpt (the highest scores per metric are in bold)

It is very interesting to highlight how the performance on “Piano” which is present in both “Classical” and “Jazz” music clips, changes greatly between the two samples. A possible explanation could be the rather more prominent role it plays in piano jazz trios, which in most cases carries the melodic part of a composition (which would explain also the high recall score). In this specific example, we see that on average the crowd workers accurately selected the small rhythmic bursts of piano play, although not as precisely. This shows that they could definitely detect its activity correctly, but could not indicate precisely its onset and offset regions.

	Accuracy	Precision	Recall
Electric Guitar	**91.7%**	96.5%	**91.6%**
Bass Guitar	82.4%	**100%**	82.4%
Drums	73%	**100%**	73%

Table 3.6: Accuracy, Precision, Recall and F-score on Rock audio excerpt (the highest scores per metric are in bold)

The participants performed better on average, in the “Rock” excerpt. We speculate that the sounds of “Electric Guitar” and “Bass Guitar” are more familiar to the demographics of the participating workers, who scored quite highly on accuracy and recall, on both instruments.

The sustained power chord of the “Electric Guitar” was easy to identify and correctly annotate its onset and offset. On the other hand, despite “Drums” and “Bass Guitar” being present during the entirety of the audio excerpt, crowd workers found “Drums” more difficult to recognize correctly, despite the results in the jazz excerpt. The difference in "Drums" between the two excerpts, shows a higher use of the snare drum in the jazz excerpt, while in the rock, the use of lower tone tom drums was more prominent.

Self-assessed Music Characteristics and Performance. In Table 3.7 we see the self-assessed "Perceptual Abilities" and self-reported "Musical Training" of the participants. The low "Musical Training" is consistent with the results of [86] but pretty low when compared to the participant pool of [65] (scoring near the bottom 30% of the population in the original study).

	Range	Median	Standard Deviation(1σ)
Perceptual Abilities	29-63	47.5	8.19
Musical Training	7-41	18.5	9.04

Table 3.7: Range, Median, Mean and Standard Deviation of Perceptual Abilities and Musical Training

The self-assessed "Perceptual Abilities" are also low compared to the sample of [65] but considerably higher than in [86]. The results in our study certainly showcase adequate perceptual skills, in regards to the task at hand.

We study the connection of their musical properties to their performance from a more qualitative perspective, due to the size of our participant pool. Their self-assessed "Perceptual Abilities" show that the users felt quite confident on the degree they can detect musical traits on sound, despite their lack of expertise as shown by their "Musical Training" average score (Table 3.7).

Comparing their assessment to their actual performance we further see that their "Musical Training" is not indicative of their capability to detect temporal activity of musical instruments. Their median score as shown on the table, is close to the low 25th percentile of the results in the original GMSI study [65], showing a generally low formal musical training. While formal training could certainly be beneficial for such tasks, people are still exposed to different musical instruments by casually enjoying music, especially as it is widely and easily accessible through streaming services. We also believe that the task design with the inclusion of an audio example of a given instrument, assisted the workers in their task to identify instruments.

Cognitive Load and Feedback. The results on the NASA-TLX questionnaire, show that from the total of 30 crowd workers, 20 found the task's difficulty average, while 17 were very confident in their performance. All of the participants reported average to low mental and physical demand, with the mental load being higher than the physical. 10 workers experienced very low temporal demand, with most finishing the study in near 10 minutes. The results though show that the workers' self-assessed performance varied greatly between individuals, with scores from "Very Low" to "Very High".

Finally, crowd workers expressed their opinions on the study through a free-form text area. Through their feedback, we found that they greatly enjoyed the study through comments such as: "*Study was very well thought out. Nothing else to add.*", "*It was fun, I would love to take part similar studies again*" and "*the study was interesting and I am* *finding the piano very interesting instrument after this study*". Some even gave their insights for future improvements in comments such as: "*Put more instruments in there*" and "*it was ok but i propose next time the sounds be played slowly for us to easily identify. thank you*".

3.3.4. DISCUSSION

Non-experts exhibited high precision with a rather high recall on most instruments, especially on the "Jazz" and "Rock" audio clips. Despite their low expertise as indicated through the "Musical Training" attribute, the results show that they were capable of perceiving the temporal activity of instruments. These abilities are in line with the findings from [86] but also people's innate understanding of music, as shown in studies [52, 45, 29].

The high precision scores combined with lower accuracy and recall scores though, could indicate that the participants underestimated the activity of the instruments in the excerpts. This means that the users although detected correctly segments of an instrument's activity, weren't able to identify the totality of temporal activity for the given instrument. By selecting more, smaller and more precise regions, one would select only the most prominent "True Positive" frames in an excerpt, but fail to select all of them, as is apparent in the cases of "Cello" and "Double Bass". Additionally, in our evaluation, we used a quite short and strict frame resolution which could potentially affect their recall scores. However, further studies are needed with variable frame resolution to test its suitability for this type of annotation task.

While it is inevitable to experience issues of sampling bias when executing crowdsourcing studies (i.e. participants will always be a smaller set of the userbase, which by itself is highly specific and smaller than the general public), we justify the differences in musical sophistication results compared to Müllensiefen et al. [65], based on the form of incentive from the side of participants to perform the study. In our case the incentive was strictly monetary, therefore we employed participants who could be less enthusiastic about music, compared to [65]. When comparing to Samiotis et al. [86] though, while the results are consistent regarding "Musical Training", the results on "Perceptual Abilities" were higher in our case, despite the use of the same crowdsourcing platform. Of course, the landscape of crowdsourcing platforms is constantly changing, but it could be a nice indication of adequately, musically perceptive crowd workers.

Finally, we believe that our interface design with the inclusion of short examples of the musical instruments on each task, must have assisted the crowd workers during annotation. We encourage further experimentation on interface design, to explore effective ways to assist workers during their audio annotation task.

Limitations. Being an exploratory study, we acknowledge that the number of participating crowd workers is lower than in traditional crowdsourcing studies. Nonetheless, we believe that the rigorous set up and the analysis of the obtained results allow us to provide valuable and robust insights, which could be used to design and deploy larger-scale studies in the future.

The music excerpts we used in our study focus on popular genres of music. As such, despite the diverse demographics of Prolific, the participants in our study were expected to be familiar with the instruments in our excerpts. We strongly encourage future studies to experiment with instruments of different traditions, as we believe that similar techniques could yield equally promising results for those instruments.

3.3.5. CONCLUSION

Our study focuses on exploring the ability of non-experts to identify the temporal activity of musical instruments in audio excerpts of western music. This is an important task during dataset production for instrument recognition, as it can provide strongly-labeled annotations which enable event detection classification tasks. In the context of classical music preservation, the non-expert participants showcased skillful recognition capabilities that can be valuable in recordings of small group composition forms of classical music. Results show that untrained crowd workers can successfully detect the activity of instruments like *clarinet* and *electric guitar*, one at a time, given an example of the instrument. The overall cognitive load that workers experienced was average, while most of them expressed their enjoyment of the tasks through free-form feedback. The positive outcomes of this work encourage conducting further studies on the topic, with focus on a larger participant pool and a more extensive evaluation dataset that includes additional genres, instruments, and identification complexities.

4
CONCLUSIONS

In this thesis, we presented our work on scaling up the digitization of classical music by incorporating the general public in the annotation of music compositions. We focused on studying the capabilities of non-experts and how to design solutions, interfaces and tasks that can be successfully used by participants of varied musical expertise, in order to expand our current music annotation practices. Through this work we diverged from the usual practices in the domain, where experts are traditionally employed to annotate data and test tools, while non-experts are usually utilized for subjective annotations of music (e.g. mood and emotions).

Our research methodology spanned from targeted experiments on UI/UX and Focus Group Discussions to larger-scale online experiments, community research, and delivering a live collaborative annotation system. We were exposed to the challenges of classical music transcription methods, where hobbyists and semi-experts have to spend considerable effort to digitally transcribe or collaboratively edit music scores.

In addition, we witnessed first-hand the lack of extensive annotated classical music data. Printing and encoding designs changed throughout centuries, leading to different formats and notations between different eras, creating a challenging landscape for automated methods. This case makes the argument regarding the popularity of composers and compositions even more relevant, as popular scores tend to have modern editions and re-prints, while compositions of more obscure composers might remain in handwritten form even now. Composers and conductors often still prefer to write music or annotate by hand, in cases accompanied by experimenting on an instrument of their choice. Only in the last decades the process has been digitized with the introduction of intuitive transcription software and audio integration, but as identified in our conversations with research partners and members of orchestras during TROMPA, composers are yet debating about best methods in writing music (by hand versus through software). Because of the late adoption of digitized and standardized methods, there are still vast amounts of corpora with varied structural and stylistic choices, resulting in automated tools for Optical Music Recognition (OMR) that are specialised in specific typesets or handwritten notations, not easily generalisable to tackle transcription holistically. These types of challenges are common when digitizing handwritten artifacts and they are still not solved at scale in classical music, resulting in the evident lack of training data and narrowly-scoped repositories of semantically rich music scores.

Scaling up these annotation processes though is a difficult task (see Section 1.3) that we approached through the design of hybrid annotation workflows. Our experimental findings showed that non-experts can meaningfully contribute to the digitization of classical music, as long as careful task and interface design have taken place. We also found that the general crowd of online crowdsourcing platforms possesses general music perception skills and sophistication that could enable use cases far beyond what is presented in this thesis. The quality of their contributions is paramount for their inclusion and the results from our targeted interface and task design experiments showcased that we can trust them with annotating complex music artifacts.

A combination of crowd work activities with that of (semi-) experts and specialized automated methods, can lead to correct, comprehensive, and varied collections of digitized music, that not only will preserve the cultural heritage of classical music but can also be used by both amateurs and professionals in the field, tackling their needs in our modern times. We strongly believe that our findings and designs, with the necessary adjustments, can be applied to other genres of music, especially to those that share similar characteristics with classical music. The skill level, sophistication, and enthusiasm that our contributors showed, certainly inspire to further examine and research their inclusion in such hybrid annotation systems.

4.1. CONTRIBUTIONS AND INSIGHTS

Our systematic analysis of the available literature and interactions with experts, helped us to identify research gaps that were vital in scaling up the classical music annotation processes. Early in this work, we understood that the hesitation to include untrained annotators in such workflows stems primarily from the complexity of music compositions and therefore the lack of trust that a non-expert could contribute valuable annotations. At the time of this thesis, the lack of previous research to showcase otherwise clearly contributed to such opinions. Transcription tasks were also often designed to tackle a score as a whole or in large portions, an admittedly daunting task even to experts, making it difficult to imagine untrained people contributing meaningfully to it. Our contributions were meant to tackle exactly such notions and shortcomings in current methods.

In the beginning, we focused on breaking down the music annotation process into smaller components, such as detecting a score's structure, transcribing notes, and evaluating the transcribed score. These components were later broken down into smaller tasks (e.g. identify clefs, input note durations), that we assessed to understand if they could be automated (based on the state-of-the-art at the time) or if they needed human annotation. The resulting tasks were later applied in small portions of the score to progressively cover it in its entirety. We quickly realised that the order of such tasks was important when iteratively transcribing a score, to preserve semantic relations between its elements. We carefully designed such tasks and delivered designs and transcription steps which we believe can act as a blueprint for future research to build upon.

Apart from breaking down a large and complex task into smaller coherent tasks, our theoretical frameworks were also useful in understanding what perceptual, comprehension, and annotation skills are needed in different stages of music transcription. Throughout our work, we contributed experiments that assess such skills, while we specifically studied how to better identify them, using psycho-acoustic and self-report methods. These computational skill-assessment methods can enable smart distribution of annotation tasks to the "right" candidate, to maximise the likelihood that an annotator is matched to a task they can perform. We also contributed User Interface (UI) designs and elements which we measured as intuitive for an untrained annotator, gathering positive feedback from both non- and semi-experts. Such results enable future researchers to re-use our designs with confidence. Meanwhile, they are presented with comprehensive evaluation methods, which can help them assess what components in their custom designs help or hinder an annotator.

Reflecting on the structure of this work and the general context and goals as described in Chapter 1, we advanced our understanding of crowd-powered annotation of music compositions in a twofold manner: a) online users showcase musical affinity and competencies that we can reliably and quantitatively assess and monitor (Chapter 2); b) careful design of annotation tasks and interfaces can enable the integration of non-experts in human-assisted music transcription systems (Chapter 3).

The outcomes of this thesis shed light on valuable aspects of how to methodologically break down annotation workflows for knowledge-intensive topics such as music transcriptions. They also underline the viability of untrained annotators and how to design targeted tasks to leverage their innate skills. Finally, they also demonstrate how and where to use automated and human annotation throughout hybrid annotation workflows for classical music. We discuss our contributions based on the research questions presented in Section 1.5 and follow their thematic segmentation across the chapters.

4.1.1. On the Musical Affinity and Competences of Non-experts

In Chapter 2, we focused on our first four Research questions:

- **[RQ1]** How could the musical profile of annotators be characterized?
- **[RQ2]** How are different music perception skills and self-reported music-related knowledge distributed among non-experts?
- **[RQ3]** How are music perception skills associated with domain and demographic attributes?
- **[RQ4]** How does the popularity of classical music composers on community-driven platforms relate to their album release trends?

While tackling **[RQ2]** we found a certain relation between specific categories of the Goldsmith's Music Sophistication Index (GMSI) [65] (namely "Perceptual Abilities" and "Musical Training") and the auditory tests of the Profile of Music Perception Skills (PROMS) [48] (namely "Melody", "Tuning", "Accent" and "Tempo"). As such, those GMSI categories should be further tested for their extent to be used as a proxy for actual perception skills, as long as the user can assess them faithfully and truthfully. When the duration of the task (and by extent, the available budget) is a concern, such an insight can help streamline the profiling stage of an annotator. An actual test of their perceptual skills is highly needed though when a researcher needs to quantitatively identify potential highly-skilled individuals. This is further encouraged by our large-scale study on Amazon Mechanical Turk

(AMT) and Prolific which showed that self-reported musical sophistication doesn't always align with the actual perceptual skills of an untrained user. Therefore, in demanding audio-based tasks, one should not rely solely on self-reported skills.

One of the biggest contributions in this chapter was regarding **[RQ3]**, where untrained, non-expert annotators of diverse demographic backgrounds, showcase an innate understanding of auditory music concepts. These skills are necessary for a plethora of audio-based tasks in the music domain, to perform knowledge-intensive tasks such as annotating the tempo of a piece, performance errors, and more. While trusting solely non-expert crowdworkers with annotating crucial parts is neither advised nor realistic, the quality of the skills in a portion of them is high enough to incorporate them in a human-assisted system. Such a system can employ concepts such as interannotator agreement to retrieve confident annotations, or profiling to assess the skills of an annotator before assigning them a task. Our design of "Musical Competence" could assist such efforts, with the results of GMSI and PROMS being "Sources" of their "Skills" and "Proficiencies", effectively tackling **[RQ1]**. In such systems, the contribution of an expert is still invaluable but their efforts can be drastically more targeted on specific portions of an annotation pipeline, such as final evaluations and annotation of highly crucial parts.

Crowdsourcing platforms though are not the only sources of potential music annotators. Our study on Wikipedia and YouTube [85] showed that online communities generate information even for the long tail of recorded canon in classical music. This shows that, with the right incentivisation mechanisms, there are online users that engage with classical music and possibly could be enthusiastic to contribute in a project involving such artifacts. A more important insight coming from that study is the possibility of finding relevant information and knowledge for classical music artifacts online. Our work on **[RQ4]**, showed that people engage with their favourite composers and compositions on online community-driven platforms. Specifically in the case of YouTube, they seemingly generate information on the long tail of popularity in a different way than Wikipedia. This could paint a picture where YouTube acts as a possible source for information gathering regarding obscure classical music artists and their works; information that otherwise would be harder (or even infeasible) to find on other sources.

4.1.2. On Deploying Human-assisted Music Transcription Systems

In Chapter 3, we focused on our technical report of our designs and implementation of our crowd-powered music transcription system, alongside our final three Research Questions in this thesis:

- **[RQ5]** To what extent can workers from microtask crowdsourcing platforms detect errors in transcribed music scores?
- **[RQ6]** To what extent can non-experts detect the onset and offset of a musical instrument's activity on polyphonic audio?
- **[RQ7]** How do their self-assessed perceptual abilities and musical knowledge relate to their performance?

Starting with our crowd-powered music transcription system, our work and resulting insights were multifaceted. This use case differed from others in this thesis as we involved the stakeholders (users interested in digital transcription) in both requirement analysis and product improvement but also assessment of transcription workflows and requirements for XML-based music score formats. Specifically, it was the only study with a targeted recruitment process, where we looked for candidates who could read and perform complex classical music (members of youth orchestras) and possibly have experience with music transcription. This was a design choice that proved invaluable for the quality of our insights in realistic digital transcription workflows that involve a group of semi-experts. Our approach for a microtask-based online transcription system, as presented in Section 3.1, was highly welcomed by the participants for its potential to alleviate the heavy and intensive workload that traditionally would be carried by one or two people in these youth orchestras.

In Chapter 3, we also presented our breakthroughs in leveraging the work of untrained users on crowdsourcing platforms for targeted transcription tasks. Such tasks could complement a transcription system, as described above, by delegating specific valuable annotation tasks to non-experts to scale up the transcription process or to generate training data for potential automated methods.

Our insights regarding **[RQ5]** in Section 3.2 show that users are capable of detecting transcription errors. Due to the limited relevant works when conducting our experiment, we specifically designed it to understand how different transcription errors and the length of the transcribed artifact could affect the crowd's performance. The crowd clearly performed better on shorter excerpts when only visually presented with the score excerpt, while the addition of audio was particularly helpful when presented with longer excerpts. Tackling **[RQ7]** in this context, showed us a strong relation between their self-reported ability to assess mistakes in performances and their actual performance in the error detection task.

We further explored the crowd's ability to annotate audio through our **[RQ6]**, in Section 3.3. We found that non-experts are quite successful in detecting not only the type of instrument performing in a given audio segment (which was already shown in [37]) but also indicate the onset and offset of its activity. Our UI design was proven helpful to the workers based on their performance, cognitive load results and feedback. Such findings are particularly encouraging for further integration of non-experts in training data creation or transcription workflows, especially for cases where we might completely lack a transcribed version of a music piece. Regarding **[RQ7]** in this context, we found results that are closer to Section 2.2: participants self-assessed their perception skills and music expertise quite low while their performance was relatively good. Although this hints to our innate ability as humans to perceive music, it also further underlines the need for proper music skill profiling other than self-assessment questionnaires, to better predict the exact capabilities of participants from an unknown pool.

4.2. CHALLENGES AND LESSONS LEARNED

Music is a complex artifact. While the melodic parts (notes, keys etc) can be a major component in compositions of certain styles, they are hardly the only parts that need to be transcribed. In a musical piece with multiple instruments, we need to understand what instruments are being used, as they need to be transcribed separately. Such a task is rather challenging, not only because the quality of the recording can affect it but also because a performer can vary an instrument's timbre during a piece, making it hard to identify it. Pauses and note durations are also important factors that are needed to identify the correct rhythm of a piece. Finally, composers use notations that are the most familiar to them and the use of standards is only encouraged when they want to "communicate" music to other composers/performers. Such challenges suggest the need for an extensive knowledge in music, to undertake the task of music transcription. This was also the main sentiment among the music experts we engaged with when discussing the design of how to include untrained annotators in transcription workflows. But music is a form of art that is widespread between cultures and one that we are exposed to from infancy (see Section 2.2 for more details). And when we strip the semantics off from music notation, identifying items, patterns and comparing portions becomes a processing task of visual/aural stimuli; a task that as humans we have exceptional skills in [101].

Use of untrained crowd. Due to the limitations of the available literature on using the crowd in music transcription, we sought to find related works on other scientific fields of crowdsourcing and cognitive sciences. We specifically studied the sophistication and perception skills of the general crowd to understand what expertise can be expected online. At the same time, we targeted small annotation tasks to examine factors that can affect their performance. As a result, while the scope of our experiments might appear very targeted, we strongly believe that the research presented here is pivotal for the future inclusion of annotators with varying expertise in music annotation. We found and showcased that in a crowd of self-identified non-experts, a large portion of individuals can perform sufficiently well in a given music task, while others can display extraordinary skills. We also showcased tasks for hybrid music annotation workflows, where participants of varying expertise can meaningfully contribute, even when they report low-to-no formal training in music. To that end, testing their perception skills is the most reliable way to model annotators' skills. However, we also found that certain thematic questions of existing established methods could approximate their music affinity, potentially reducing the costs of profiling non-recurrent users. These are breakthrough findings that encourage future crowdsourcing research on music transcription, as our systematic approach in evaluating the performance and skills of the crowd enables the generalisability of our findings in the field.

Enthusiasm about music. Our access to human participants was diverse and inclusive. Contrary to previous works in the domain where human transcription was involved, we conducted research with a wide variety of participants, employing a total of 374 crowd workers, recruiting 64 semi-experts from youth orchestras in the Netherlands, and analysing community-driven information for nearly 6,000 classical music composers on both YouTube and Wikipedia. Feedback from active participants was very encouraging: while people on online platforms liked our tasks and wanted to return in future studies, semi-experts found our approach relaxing, as micro transcription tasks alleviated the "tediousness" of transcribing a music score. On the other hand, participants already accustomed to transcribing music wanted to work with longer excerpts. The small portions of the score that were presented to them, were breaking the themes and motifs of a piece, making the task more "mechanical" for those annotators. For the untrained crowd, though, the tasks only caused low to medium cognitive load, with the inclusion of audio excerpts being a pleasant break between more traditional annotation tasks. Alongside the enthusiasm that users showcase when sharing music-related information on community-driven platforms, we witnessed that people in general can be self-motivated when it comes to generating information for music. Such enthusiasm should encourage future works on the crowd's inclusion, as the music domain is in need of diverse, openly available annotated datasets and we believe that non-experts can play a crucial role in scaling up those processes.

Handwritten musical notes. The corpus of classical music is immense, covering an ever-expanding set of composers spanning centuries throughout history. The "classical music" genre covers an extensive set of musical styles, each with its own musical structures and instrumentation. Despite current notation practices being fairly standardized, handwritten transcriptions, overlayed corrections, or adjustments on the corners of pages are very real and valid challenges.

Through our interaction with orchestras, we found that such information is very valuable to performers and music librarians. Notations and alterations by famous conductors are invaluable to youth orchestras, who argued that knowing how other orchestras changed famous pieces could inform their versions. Also, large orchestral organizations like The Royal Concertgebouw Orchestra in Amsterdam, want to preserve alterations made by famous conductors, for future reference or historical reasons. Such notations though are notoriously difficult to tackle (see Section 1.3). They could span from textual notes to symbols overlayed on top of existing ones to whole systems of notes altering completely a passage. While we did not tackle such challenges in this work, we believe they need special attention in the future because of the value and enrichment they bring into existing music transcriptions.

Transcription tooling. When discussing preserving music transcriptions, we also needed to identify existing tools and the format in which we would store our transcriptions. We quickly found that XML-based documents are the most popular digital format for semantically rich music scores, while most robust transcription tools were proprietary. Although graph representations of music scores have been examined in the literature [23, 7, 72], XML documents are an established "tradition" in digital semantic representations of music scores, and it was the chosen format to store our transcriptions in this thesis. We discovered certain challenges that one needs to be mindful of when designing transcription microtasks using XML as the target format. When working with MEI in this thesis, we found that a certain pre-processing of the segments is required, as the task needs to show only portions of the score, but the rendering module needs information regarding the whole XML document. Certain score information such as key and cleffs, need to be prioritised properly as they will affect all subsequent music information. This will ensure the correct rendering of the score segments to the annotators, while the system processes the document. Additionally, when following the microtask design workflow, one should assign more than one annotator per segment-annotation tuple and measure the inter-annotator agreement to resolve any conflicts. In the case of score transcription, the rendering of an annotated score segment might look the same between two annotators, but the underlining, generated XML snippet might differ. This creates a challenge that a researcher might have to tackle effectively through tree comparison algorithms.

Our UI designs and tools were inspired by popular existing tools but adapted to our purpose. Throughout the thesis, we assumed that, to an untrained crowd, music notation would already be novel and "distracting", and we, therefore, opted for minimalism and familiarity as preferable design choices. Additionally, sound proved a strong stimulus to the participants, especially when complex and longer excerpts needed to be annotated. This is a result that is in line with the experimental findings of the “CosmoNote” system [24], created and published at the end of this PhD trajectory, which provides long audio excerpts for performance annotation using a Web-based system for citizen science.

In our studies, the feedback from both non-experts and the participants from youth orchestras was encouraging and we learned a lot about human-computer interaction in the context of music transcription. Contrarily to our assumption, we found that untrained workers were not overwhelmed by music notation, and found our studies “well thought out”, while the included audio excerpts were perceived as “fun” and engaging. In all of our studies where audio was included, we also found that crowd workers possessed audio equipment of adequate quality, a fact that could have resulted by the COVID-19 safety measures and the rise of remote working.

Despite publishing our solutions, we have to point out the general lack of proper tooling that can enable researchers of different technical backgrounds to pursue the principles of this thesis. Tools and interfaces that are familiar to (semi-)experts are often proprietary, where researchers need to acquire licenses to use them. Especially in the context of microtask design and online hosting of the tasks, these tools prove to be limited, as they are not designed for such cases. This leaves only a few choices of freely available and customisable digital score transcription tools, which are often tailored to specific output formats. For the domain to flourish, we need coordination among the research community to develop standards and tools that can enable researchers to freely experiment with hybrid-annotation workflows for music.

4.3. Future Opportunities

Through this thesis, we wanted to research the capabilities of untrained annotators and push past the general notions regarding their skills and performance in knowledge-demanding tasks. Their inclusion is pivotal for the success of large-scale, hybrid intelligence systems that specialise in music digitization and annotation. They can assist in expanding our efforts to create training data, support evaluation processes, and aid experts by contributing to laborious tasks. Our findings on classical music have already been useful in studies of other music traditions [58, 105] and we believe they can further inform future research for including untrained contributors in music annotation processes.

Through this work, we also obtained insights that apply beyond the scope of this thesis. Firstly, when looking into music comprehension and cognitive mechanisms of people, we were exposed to literature on biases that might be present in annotators, which can affect their performance in music tasks. Enculturation has shown to play a certain role in musical memory and performance evaluation tasks [22, 4], while the emotional response has shown to be more universal [3], with the phenomenon only affecting the perception of more specific music styles [8]. From our findings and previous research [56] though, we believe that properly exposing annotators to examples of unfamiliar musical concepts can minimise the effects of such biases.

This could be explained by common characteristics that are shared between music traditions. In the case of classical music, repetition of parts, their reintroduction and expansion on them, is a prevalent characteristic of the genre but not unique to it. In Jazz and Traditional/Folk music, composers often utilize such creative methods, while referencing other music pieces or borrowing from other composers is also prevalent and often used as a "homage" to other composers or compositions. Other music genres are also well known to have repetitive motifs and well-defined structures, such as "Schlager" music (European-style of popular music) and the traditional Christmas and children's songs. Such musical characteristics could potentially be utilized as "anchors" to familiarise untrained annotators with music traditions outside their experiences.

Finally, focusing specifically on preserving classical music information, we need to be mindful of the plurality, multiplicity, and range of the tradition. Classical music composers would be inspired by the tradition of their contemporary era, but often combine elements uniquely while pushing the boundaries of the tradition to "invent" their own style. In a comprehensive score analysis, we should be able to store such information, by identifying its parts and the recognized traditions it follows. We should also connect the music piece and its composer to potentially other composers who rearranged the original score or from whom the original composer might have borrowed. Through our interactions with experts, we found that tracking such connections is very important for preserving the history and the heritage of classical music. This can prove to be a particularly challenging feat in a tradition of such longevity, but we believe that the general crowd has showcased perception, enthusiasm, and attention skills that could be leveraged to embark on such attempts. Works like those presented in this thesis, can provide proper tooling and task designs that can assist both the crowd in their work, but also enable their "co-existence" alongside experts and automated methods, making such ambitious goals more attainable.

BIBLIOGRAPHY

[1] Moshe Adler. Stardom and talent. *The American economic review*, 75(1):208–212, 1985.

[2] Moshe Adler. Stardom and talent. *Handbook of the Economics of Art and Culture*, 1:895–906, 2006.

[3] Heike Argstatter. Perception of basic emotions in music: Culture-specific or multicultural? *Psychology of Music*, 44(4):674–690, 2016.

[4] Gökhan Aydogan, Nicole Flaig, Srekar N Ravi, Edward W Large, Samuel M McClure, and Elizabeth Hellmuth Margulis. Overcoming bias: Cognitive control reduces susceptibility to framing effects in evaluating musical performance. *Scientific reports*, 8(1):6229, 2018.

[5] Siamak Baharloo, Paul A Johnston, Susan K Service, Jane Gitschier, and Nelson B Freimer. Absolute pitch: an approach for identification of genetic and nongenetic components. *The American Journal of Human Genetics*, 62(2):224–231, 1998.

[6] Siamak Baharloo, Susan K Service, Neil Risch, Jane Gitschier, and Nelson B Freimer. Familial aggregation of absolute pitch. *The American Journal of Human Genetics*, 67(3):755–758, 2000.

[7] David Bainbridge and Tim Bell. A music notation construction engine for optical music recognition. *Software: Practice and Experience*, 33(2):173–200, 2003.

[8] Laura-Lee Balkwill, William Forde Thompson, and RIE Matsunaga. Recognition of emotion in japanese, western, and hindustani music by japanese listeners 1. *Japanese Psychological Research*, 46(4):337–349, 2004.

[9] Elizabeth Francis Beach, Warwick Williams, and Megan Gilliver. The objective-subjective assessment of noise: Young adults can estimate loudness of events and lifestyle noise. *International journal of audiology*, 51(6):444–449, 2012.

[10] Alejandro Bellogín, Arjen P de Vries, and Jiyin He. Artist popularity: Do web and social music services agree? In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[11] Juan J Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera. A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *ISMIR*, pages 559–564. Citeseer, 2012.

[12] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Andrea Mauri. Reactive crowdsourcing. In *Proceedings of the 22nd international conference on World Wide Web*, pages 153–164, 2013.

[13] Oliver Budzinski and Sophia Gaenssle. The economics of social media stars: an empirical investigation of stardom, popularity, and success on youtube. *Ilmenau Economics Discussion Papers*, 21(112), 2018.

[14] Manuel Burghardt and Sebastian Spanner. Allegro: User-centered design of a tool for the crowdsourced transcription of handwritten music scores. In *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, pages 15–20, New York, NY, USA, 2017. ACM.

[15] Anja Burkhard, Stefan Elmer, and Lutz Jäncke. Early tone categorization in absolute pitch musicians is subserved by the right-sided perisylvian brain. *Scientific reports*, 9(1):1–14, 2019.

[16] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P Bello, and Oded Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21, 2017.

[17] Christopher Cayari. The youtube effect: How youtube has provided new ways to consume, create, and share music. *International Journal of Education & the Arts*, 12(6):n6, 2011.

[18] Gloria Chatzopoulou, Cheng Sheng, and Michalis Faloutsos. A first step towards understanding popularity in youtube. In *2010 INFOCOM IEEE Conference on Computer Communications Workshops*, pages 1–6. IEEE, 2010.

[19] Graham Cheetham and Geoffrey E Chivers. *Professions, competence and informal learning*. Edward Elgar Publishing, 2005.

[20] Liang Chen and Christopher Raphael. Human-Directed Optical Music Recognition. *Electronic Imaging*, 2016(17):1–9, 02 2017.

[21] Pauline Degrave and Jonathan Dedonder. A french translation of the goldsmiths musical sophistication index, an instrument to assess self-reported musical skills, abilities and behaviours. *Journal of New Music Research*, 48(2):138–144, 2019.

[22] Steven M Demorest, Steven J Morrison, Denise Jungbluth, and Münir N Beken. Lost in translation: An enculturation effect in music memory performance. *Music Perception*, 25(3):213–223, 2008.

[23] Hoda Fahmy and Dorothea Blostein. A graph grammar programming style for recognition of music notation. *Machine Vision and Applications*, 6:83–99, 1993.

[24] Lawrence Fyfe, Daniel Bedoya, and Elaine Chew. Annotation and analysis of recorded piano performances on the web. *Journal of the Audio Engineering Society*, 70(11):962–978, 2022.

[25] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–29, 2017.

[26] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, pages 6–26. Springer, 2017.

[27] Basilios Gatos, Ioannis Pratikakis, and Stavros J Perantonis. An adaptive binarization technique for low quality historical documents. In *International Workshop on Document Analysis Systems*, pages 102–113. Springer, 2004.

[28] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[29] Bruno Gingras, Henkjan Honing, Isabelle Peretz, Laurel J Trainor, and Simon E Fisher. Defining the biological bases of individual differences in musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140092, 2015.

[30] Mark Gotham, Peter Jonas, Bruno Bower, William Bosworth, Daniel Rootham, and Leigh VanHandel. Scores of scores: an openscore project to encode and share sheet music. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, pages 87–95, 2018.

[31] Rebecca A. Grier. How high is high? a meta-analysis of nasa-tlx global workload scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1):1727–1731, 2015.

[32] Sascha Grollmisch, Estefanía Cano, Christian Kehling, and Michael Taenzer. Analyzing the potential of pre-trained embeddings for audio classification tasks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 790–794. IEEE, 2021.

[33] Yoonchang Han, Jaehun Kim, and Kyogu Lee. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):208–221, 2016.

[34] Erin E Hannon and Laurel J Trainor. Music acquisition: effects of enculturation and formal training on development. *Trends in cognitive sciences*, 11(11):466–472, 2007.

[35] Zofia Hanusz, Joanna Tarasinska, and Wojciech Zielinski. Shapiro-wilk test with known mean. *REVSTAT-Statistical Journal*, 14(1):89–100, 2016.

[36] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[37] Eric Humphrey, Simon Durand, and Brian McFee. Openmic-2018: An open data-set for multiple instrument recognition. In *ISMIR*, pages 438–444, 2018.

[38] Krista L Hyde and Isabelle Peretz. Brains that are out of tune but in time. *Psychological science*, 15(5):356–360, 2004.

[39] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

[40] Graham Jones, Bee Ong, Ivan Bruno, and NG Kia. Optical music imaging: music document digitisation, recognition, evaluation, and restoration. In *Interactive multimedia music technologies*, pages 50–79. IGI Global, 2008.

[41] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396, 2015.

[42] L Juri, Eelco Herder, Arne Koesling, Christoph Lofi, Daniel Olmedilla, Odysseas Papapetrou, and Wolf Siberski. A model for competence gap analysis. In *Proceedings of 3rd International Conference in WEB Information Systems and technology Barcelona, Spain*. Citeseer, 2007.

[43] Simon Kassing, Jasper Oosterman, Alessandro Bozzon, and Geert-Jan Houben. Locating domain-specific contents and experts on social bookmarking communities. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, page 747–752, New York, NY, USA, 2015. Association for Computing Machinery.

[44] Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps. *EURASIP Journal on Advances in Signal Processing*, 2007:1–15, 2006.

[45] Stefan Koelsch, Katrin Schulze, Daniela Sammler, Thomas Fritz, Karsten Müller, and Oliver Gruber. Functional architecture of verbal and tonal working memory: an fmri study. *Human brain mapping*, 30(3):859–873, 2009.

[46] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[47] Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. Tagatune: A game for music and sound annotation. In *ISMIR*, volume 3, page 2, 2007.

[48] Lily NC Law and Marcel Zentner. Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PloS one*, 7(12):e52508, 2012.

[49] Jin Ha Lee. Crowdsourcing music similarity judgments using mechanical turk. In *ISMIR*, pages 183–188, 2010.

[50] Jin Ha Lee, Trent Hill, and Lauren Work. What does music mood mean for real users? In *Proceedings of the 2012 IConference*, iConference '12, page 112–119, New York, NY, USA, 2012. Association for Computing Machinery.

[51] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 129–138, 2012.

[52] Alvin M Liberman and Ignatius G Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

[53] César F Lima, Ana Isabel Correia, Daniel Müllensiefen, and São Luís Castro. Goldsmiths musical sophistication index (gold-msi): Portuguese version and associations with socio-demographic factors, personality and music preferences. *Psychology of Music*, 48(3):376–388, 2020.

[54] Hsin-Rui Lin, Reinhard Kopiez, Daniel Müllensiefen, and Anna Wolf. The chinese version of the gold-msi: Adaptation and validation of an inventory for the measurement of musical sophistication in a taiwanese sample. *Musicae Scientiae*, 25(2):226–251, 2021.

[55] Christoph Lofi and Kinda El Maarry. Design patterns for hybrid algorithmic-crowdsourcing workflows. In *2014 IEEE 16th Conference on Business Informatics*, volume 1, pages 1–8. IEEE, 2014.

[56] Psyche Loui, David L Wessel, and Carla L Hudson Kam. Humans rapidly learn grammatical structure in a new musical scale. *Music perception*, 27(5):377–388, 2010.

[57] Glenn M MacDonald. The economics of rising stars. *The American Economic Review*, pages 155–166, 1988.

[58] Yonatan Malin, Christina Crowder, Clara Byom, and Daniel Shanahan. Community based music information retrieval: A case study of digitizing historical klezmer manuscripts from kyiv. *Transactions of the International Society for Music Information Retrieval*, 5(1):208–221, 2022.

[59] Michael I Mandel, Douglas Eck, and Yoshua Bengio. Learning tags that vary within a song. *ISMIR, Utrecht, Netherlands*, 2010.

[60] Kelsey Mankel and Gavin M Bidelman. Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proceedings of the National Academy of Sciences*, 115(51):13129–13134, 2018.

[61] Matija Marolt, Ciril Bohak, Alenka Kavčič, and Matevž Pesek. Automatic segmentation of ethnomusicological field recordings. *Applied Sciences*, 9(3):439, 2019.

[62] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Chatterbox: Conversational interfaces for microtask crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 243–251, 2019.

[63] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016.

[64] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, 8(8), 2013.

[65] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, 9(2):e89642, 2014.

[66] Leann Myers and Maria J Sirois. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12, 2004.

[67] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566, 2010.

[68] Jieun Oh and Ge Wang. Evaluating crowdsourcing through amazon mechanical turk as a technique for conducting music perception experiments. In *Proceedings of the 12th International Conference on Music Perception and Cognition*, pages 1–6, 2012.

[69] Jasper Oosterman, Alessandro Bozzon, Geert-Jan Houben, Archana Nottamkandath, Chris Dijkshoorn, Lora Aroyo, Mieke HR Leyssen, and Myriam C Traub. Crowd vs. experts: nichesourcing for knowledge intensive tasks in cultural heritage. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 567–568, 2014.

[70] Jasper Oosterman, Jie Yang, Alessandro Bozzon, Lora Aroyo, and Geert-Jan Houben. On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks. *Computer Networks*, 90:133 – 149, 2015. Crowdsourcing.

[71] Claudia Orellana-Rodriguez, Ernesto Diaz-Aviles, and Wolfgang Nejdl. Mining emotions in short films: user comments or crowdsourcing? In *Proceedings of the 22nd International Conference on World Wide Web*, pages 69–70. ACM, 2013.

[72] Alexander Pacha, Jorge Calvo-Zaragoza, and Jan Hajic Jr. Learning notation graph construction for full-pipeline optical music recognition. In *ISMIR*, pages 75–82, 2019.

[73] John Paolillo, Sharad Ghule, and Brian Harper. A network view of social media platform history: Social structure, dynamics and content on youtube. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[74] Isabelle Peretz and Krista L Hyde. What is specific to music processing? insights from congenital amusia. *Trends in cognitive sciences*, 7(8):362–367, 2003.

[75] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2744–2748. IEEE, 2017.

[76] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, Andre RS Marcal, Carlos Guedes, and Jaime S Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.

[77] ITU Recommendation. General methods for the subjective assessment of sound quality. *ITU-R BS*, pages 1284–1, 2003.

[78] Lindsey Reymore and Niels Chr Hansen. A theory of instrument-specific absolute pitch. *Frontiers in psychology*, 11:2801, 2020.

[79] Florence Rossant and Isabelle Bloch. Robust and adaptive omr system including fuzzy modeling, fusion of musical rules, and possible error detection. *EURASIP Journal on Advances in Signal Processing*, 2007:1–25, 2006.

[80] Charalampos Saitis, Andrew Hankinson, and Ichiro Fujinaga. Correcting large-scale omr data with crowdsourcing. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–3, 2014.

[81] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 344–348. IEEE, 2017.

[82] Ioannis Petros Samiotis, Christoph Lofi, Shaad Alaka, Cynthia CS Liem, and Alessandro Bozzon. Scriptoria: A crowd-powered music transcription system. In *Companion Proceedings of the Web Conference 2022*, pages 256–259, 2022.

[83] Ioannis Petros Samiotis, Christoph Lofi, and Alessandro Bozzon. Hybrid annotation systems for music transcription. In *3rd International Workshop on Reading Music Systems*, 2021.

[84] Ioannis Petros Samiotis, Christoph Lofi, and Alessandro Bozzon. Crowd's performance on temporal activity detection of musical instruments in polyphonic music. In Augusto Sarti, Fabio Antonacci, Mark Sandler, Paolo Bestagini, Simon Dixon, Beici Liang, Gaël Richard, and Johan Pauwels, editors, *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, pages 612–618, 2023.

[85] Ioannis Petros Samiotis, Andrea Mauri, Chirstoph Lofi, and Alessandro Bozzon. On the popularity of classical music composers on community-driven platforms. In *International Conference on Web Engineering*, pages 327–335. Springer, 2023.

[86] Ioannis Petros Samiotis, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, and Alessandro Bozzon. Exploring the music perception skills of crowd workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 108–119, 2021.

[87] Ioannis Petros Samiotis, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, and Alessandro Bozzon. An analysis of music perception skills on crowdsourcing platforms. *Frontiers in Artificial Intelligence*, 5, 2022.

[88] Ioannis Petros Samiotis, Sihang Qiu, Andrea Mauri, Cynthia C. S. Liem, Christoph Lofi, and Alessandro Bozzon. Microtask crowdsourcing for music score transcriptions: An experiment with error detection. In Julie Cumming, Jin Ha Lee, Brian McFee, Markus Schedl, Johanna Devaney, Cory McKay, Eva Zangerle, and Timothy de Reuse, editors, *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, pages 901–907, 2020.

[89] Neela Sawant, Jia Li, and James Z Wang. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools and Applications*, 51(1):213–246, 2011.

[90] Nora K Schaal, Anna-Katharina R Bauer, and Daniel Müllensiefen. Der gold-msi: replikation und validierung eines fragebogeninstrumentes zur messung musikalischer erfahrenheit anhand einer deutschen stichprobe. *Musicae Scientiae*, 18(4):423–447, 2014.

[91] Markus Schedl and Marko Tkalčič. Genre-based analysis of social media data on music listening behavior: are fans of classical music really averse to social media? In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, pages 9–13, 2014.

[92] Hendrik Schreiber and Meinard Müller. A crowdsourced experiment for tempo estimation of electronic dance music. In *ISMIR*, pages 409–415, 2018.

[93] Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. Opinion mining on YouTube. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1252–1261, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[94] Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. Multi-lingual opinion mining on youtube. *Information Processing & Management*, 52(1):46–60, 2016.

[95] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, pages 891–900. ACM, 2010.

[96] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)*, 8(3):17, 2014.

[97] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE, 2008.

[98] Jacquelin A Speck, Erik M Schmidt, Brandon G Morton, and Youngmoo E Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *ISMIR*, volume 104, pages 549–554. Citeseer, 2011.

[99] George J Stigler and Gary S Becker. De gustibus non est disputandum. *The american economic review*, 67(2):76–90, 1977.

[100] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[101] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

[102] PA Totterdell and Karen Niven. *Workplace moods and emotions: A review of research.* Createspace Independent Publishing, 2014.

[103] Fredrik Ullén, Miriam A Mosing, Linus Holm, Helene Eriksson, and Guy Madison. Psychometric properties and heritability of a new online test for musicality, the swedish musical discrimination test. *Personality and Individual Differences*, 63:87–93, 2014.

[104] Julián Urbano, Jorge Morato, Mónica Marrero, and Diego Martín. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *ACM SIGIR workshop on crowdsourcing for search evaluation*, pages 9–16. ACM New York, 2010.

[105] Laura Vázquez-Fragua, Miguel Ángel Fernández-Blázquez, and José María Ruiz-Sánchez de León. Validation of the abbreviated version of the profile of musical perception skills (mini-proms) in a spanish sample of musicians and non-musicians. *Musicae Scientiae*, page 10298649241227812, 2024.

[106] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. The youtube social network. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[107] Pengyu Wei and Ning Wang. Wikipedia and stock return: Wikipedia usage pattern helps to predict the individual stock movement. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 591–594, 2016.

[108] Paul D Werner, Alan J Swope, and Frederick J Heide. The music experience questionnaire: Development and correlates. *The Journal of psychology*, 140(4):329–345, 2006.

[109] David L Wessel. Timbre space as a musical control structure. *Computer music journal*, pages 45–52, 1979.

[110] Tingxin Yan, Vikas Kumar, and Deepak Ganesan. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 77–90, 2010.

[111] Marcel Zentner and Hannah Strauss. Assessing musical ability quickly and objectively: development and validation of the short-proms and the mini-proms. *Annals of the New York Academy of Sciences*, 1400(1):33–45, 2017.

[112] Mengdie Zhuang and Ujwal Gadiraju. In what mood are you today? an analysis of crowd workers' mood, performance and engagement. In *Proceedings of the 10th ACM Conference on Web Science*, pages 373–382, 2019.

LIST OF FIGURES

1.1 A simplified example schema of music concepts related to a classical music composition. “Composition” is the musical work of a “Composer”, published in physical media such as a “Score” (or “sheet music”). A “Score” can be edited (by an “Editor”) and re-published, while there can be multiple audio/video recordings of a musical work performed on specific “Instruments” by specific “Performers”. 2

1.2 A starting system of measures in a piano piece (Andante, in F major, Ludwig van Beethoven) 5

1.3 Examples of handwritten music scores. 8

1.4 An example of handwritten notes overlayed on top of a printed score (Mahler 4th Symphony) 9

1.5 High-level Hybrid Processing Workflows 10

2.1 Modeling annotator competence and task requirements 17

2.2 The four steps in the music perception skills study. 22

2.3 Interfaces of the study (a, GMSI questionnaire, b, Mini-PROMS, c, Participant ID prompt. 26

2.4 Musical Training (GMSI) and Performance on Mini-PROMS (acc%). 37

2.5 Jaccard similarity at different ranking cut-offs with normal (a) and reverse ranking (b) 46

3.1 Schema of Crowd Task Manager’s modules. 51

3.2 Improved designs for the transcription tasks. 58

3.3 Microtask User Interfaces: (a) Visual, (b) Audio and (c) Combination . . . 62

3.4 Workers error detection accuracy (unit:%) (a) per user interface and (b) per segment size. 65

3.5 Aggregated error detection accuracy (unit:%) (a) per user interface and (b) per segment size. 66

3.6 Worker execution time (unit: seconds) of each microtask by (a) user interface and (b) segment size. 67

3.7 Main audio annotation task 71

3.8 Task instructions and warning 73

LIST OF TABLES

1 Summary of key terms and corresponding definitions. xi

2.1 Prolific participant demographics 28
2.2 GMSI Range, Median, Mean and Standard Deviation 29
2.3 Intercorrelations (Pearson's R) of subscales of GMSI scores. 29
2.4 Mini-PROMS Range, Median, Mean and Standard Deviation 30
2.5 MTurk participant demographics 33
2.6 GMSI Range, Median, Mean and Standard Deviation 34
2.7 Intercorrelations (Spearman's Rank) of subscales of GMSI scores. 34
2.8 Mini-PROMS Range, Median, Mean and Standard Deviation 35
2.9 Spearman correlation the metrics of Wikipedia and Youtube 45
2.10 Spearman correlation between the platform rankings and album releases. 45

3.1 Precision, Recall and Accuracy of *individual* answers, by segment sizes and interfaces. 64
3.2 Overall Precision, Recall and Accuracy of *aggregated* answers by segment sizes and interfaces. In bold you find the highest precision, recall and accuracy by segment size, while underlined you find the highest overall performance 64
3.3 Participant demographics 74
3.4 Accuracy, Precision, Recall and F-score on Classical audio excerpt (the highest scores per metric are in bold) 74
3.5 Accuracy, Precision, Recall and F-score on Jazz audio excerpt (the highest scores per metric are in bold) 75
3.6 Accuracy, Precision, Recall and F-score on Rock audio excerpt (the highest scores per metric are in bold) 75
3.7 Range, Median, Mean and Standard Deviation of Perceptual Abilities and Musical Training 76

CURRICULUM VITÆ

Ioannis Petros SAMIOTIS

18-05-1989 Born in Athens, Greece.

EDUCATION

2018–2024
Ph.D. in Human Computation and Music Information
Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Department of Software Technology
Web Information Systems group
Netherlands
Thesis: Crowd-Assisted Annotation
of Classical Music Compositions
Promotor: Prof. dr. ir. G.J.P.M. Houben
Co-Promotor: Prof. dr. ir. A. Bozzon
*Daily Supervisor:*Dr. C. Lofi

2015–2018
M.Sc. in Computer Science
Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Netherlands
Thesis: Side-Channel Attacks
using Convolutional Neural Networks
Promotor: Dr. S. Picek

2011–2012
M.Sc. in Digital Visual Effects
University of Kent
Faculty of Engineering and Digital Arts
United Kingdom
Thesis: The Mirror (Short Film)
Promotor: D. Byers Brown

2007–2011
B.Sc. in Cultural Technology and Communication
University of Aegean
Faculty of Social Sciences
Greece

1995–2009	Piano and Music Theory Studies Nikos Skalkotas Conservatory Greece

Professional Experience

2022–2024	Data Scientist, XITE Amsterdam, Netherlands
2021	Audio/Video Production of Short-form Documentary, Center of Urban Studies Amsterdam, Netherlands
2017	Research Internship, Riscure Delft, Netherlands
2014–2015	Collaborating Researcher, University of Aegean Mytilene, Greece

LIST OF PUBLICATIONS

9. 🗎 **Ioannis Petros Samiotis**, Christoph Lofi, and Alessandro Bozzon. Crowd's performance on temporal activity detection of musical instruments in polyphonic music, Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023, pages 612–618, 2023.

8. 🗎 **Ioannis Petros Samiotis**, Andrea Mauri, Chirstoph Lofi, and Alessandro Bozzon. On the popularity of classical music composers on community-driven platforms. In International Conference on Web Engineering, pages 327–335. Springer, 2023.

7. 🗎 **Ioannis Petros Samiotis**, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, and Alessandro Bozzon. An analysis of music perception skills on crowdsourcing platforms. Frontiers in Artificial Intelligence, 5, 2022.

6. 🗎 **Ioannis Petros Samiotis**, Christoph Lofi, Shaad Alaka, Cynthia CS Liem, and Alessandro Bozzon. Scriptoria: A crowd-powered music transcription system. In Companion Proceedings of the Web Conference 2022, pages 256-259, 2022.

5. 🗎 **Ioannis Petros Samiotis**, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, and Alessandro Bozzon. Exploring the music perception skills of crowd workers. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 9, pages 108–119, 2021

4. 🗎 **Ioannis Petros Samiotis**, Christoph Lofi, and Alessandro Bozzon. Hybrid annotation systems for music transcription. In 3rd International Workshop on Reading Music Systems, 2021.

3. 🗎 **Ioannis Petros Samiotis**, Sihang Qiu, Andrea Mauri, Cynthia C. S. Liem, Christoph Lofi, and Alessandro Bozzon. Microtask crowdsourcing for music score transcriptions: An experiment with error detection, Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020, pages 901–907, 2020

2. Stjepan Picek, **Ioannis Petros Samiotis**, Jaehun Kim, Annelie Heuser, Shivam Bhasin, and Axel Legay. On the performance of convolutional neural networks for side-channel analysis. In Security, Privacy, and Applied Cryptography Engineering: 8th International Conference, SPACE 2018, Kanpur, India, December 15-19, 2018, Proceedings 8, pp. 157-176. Springer International Publishing, 2018.

1. Hammudoglu, J. S., J. Sparreboom, J. I. Rauhamaa, J. K. Faber, L. C. Guerchi, **Ioannis Petros Samiotis**, S. P. Rao, and Johan A. Pouwelse. Portable trust: biometric-based authentication and blockchain storage for self-sovereign identity systems. arXiv preprint arXiv:1706.03744 (2017).

🗎 Included in this thesis.

SIKS Dissertation Series

Since 1998, all dissertations written by Ph.D. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series.

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (TiU), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VUA), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TU/e), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VUA), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description

24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

---

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems

08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks

10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science

11 withdrawn

12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning

13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling

15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health

16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems

17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design

18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations

19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior

20 Ritsart Anne Plantenga (UL), Omgang met Regels

21 Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour

25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics

26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework

27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases

30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting

31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows

32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining

33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method

35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach

36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs

37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models

38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings

39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance

40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models

41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines

42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis

43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

44 Marcel Schmitz (OU), Towards learning analytics-supported learning design

45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology

47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs

48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

---

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems

02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks